

Demographic and Psychographic Estimation of Twitter Users Using Social Structures

Jun Ito, Kyosuke Nishida, Takahide Hoshide, Hiroyuki Toda
and Tadasu Uchiyama

Abstract Word-of-mouth marketing on social media has become more urgent with the increasing number of users and posts, and it is important to estimate user attributes because most users on Twitter do not reveal their attributes. We propose new methods 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. This study has three contributions on the task of user attribute estimation. First, we investigate a labeling method that finds the users associated with a blog account and uses their profile attributes on blog as true labels of training tweet data. We confirm that using the blog labels achieved higher accuracy than manual labeling and pattern matching methods, with respect to four attributes (gender, age, occupation, and interests). Second, we validate the best way to combine bag-of-words features of profile documents and tweets. We evaluate nine combining methods and show that words in profile documents should be treated distinctively from those in tweets. Third, we reveal that to adjust amount of information from social neighbors affects estimation accuracy. We experiment three adjustment levels and show that our method, which utilizes the target user's profile document and tweets and the neighbors' profile documents (not including tweets), achieved the best accuracy. Overall experiments conducted on the estimation of the four attributes show that our method achieved higher accuracy than conventional methods that use manually-labeled tweets.

Keywords User attribute estimation · Twitter · Microblog · Social media · Blog

J. Ito (✉), T. Hoshide, H. Toda · T. Uchiyama
NTT Service Evolution Laboratories, NTT Corporation, 1-1 Hikari-no-oka,
Yokosuka-shi, Kanagawa 239-0847, Japan
e-mail: ito.jun@lab.ntt.co.jp

K. Nishida
NTT Resonant Inc, 4-1-8F Granpark Tower, Shibaura 3-chome Minato-ku,
Tokyo 108-0023, Japan
e-mail: k-nishi@ntr.co.jp

1 Introduction

Social media represented by Facebook¹ and Twitter² has grown rapidly in the last few years. The number of monthly active users on Facebook was more than one billion as of October 2012,³ while Twitter had more than 200 million as of March 2013.⁴ In addition, Facebook had 3.2 billion likes and comments per day and 300 million new photos daily in March 2012,⁵ and Twitter had 400 million tweets per day in March 2013 (See footnote 4). Social media has attracted an unbelievable number of users who post many of their daily actions, events, and communication exchanges.

Word-of-mouth marketing on social media has become crucial with the increasing number of users and posts [9, 16]. Many opinions and impressions about products or services are among the social media posts. So social media marketing is attracting a lot of attention. The most common way to understand consumers' opinions and impressions about products or services is a questionnaire survey. However, monetary costs can be high depending on the number of monitors and questions, and real time response is not possible because it takes a long time to complete the investigation and summarize the findings. On the other hand, word-of-mouth marketing on social media can investigate many opinions and impressions at low monetary cost and in real time. Because of these benefits, many companies are working on word-of-mouth marketing tools on social media, and various marketing applications such as Radian6, Sysomos, and Foresight, are on the market.

Although word-of-mouth marketing on social media has benefits in terms of monetary cost and rapid response, its key disadvantage is that it is hard to acquire user attributes. Opinions and impressions vary according to both demographic (gender, age, occupation, etc.) and psychographic (interests, etc.) attributes, so marketers analyze opinions and impressions for each attribute or examine the distribution of attributes. Questionnaire surveys can specifically ask for the user attributes, but this is not easy with social media because most users do not specify their attributes in the user profile description.

To resolve this issue, we address user attribute estimation on Twitter, which has many users and posts and the data are publicly available. We estimate four attributes (gender, age, occupation, and interests) of a Twitter user from the contents (a profile document and tweets) generated by the user and the user's social neighbors, i.e. those with whom the user has conversed (mentioned).

Contributions. (i) We investigate a labeling method that finds the users associated with a blog account and uses their profile attributes on blog as true labels of training

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/>.

tweet data. This idea was conceived by Burger et al. [2], and they demonstrated the effectiveness in gender attribute and investigated the validity by using 1,000 random sampled users. On the other hand, we confirm that using the blog labels achieved higher accuracy than manual labeling and pattern matching methods, with respect to four attributes (gender, age, occupation, and interests). (ii) We validate the best way to combine bag-of-words features of profile documents and tweets. Profile document is a high quality source of information since it may hold actual user attributes, so estimation accuracy can be increased by using it together with tweets as the training data. We evaluate nine combining methods and show that words in profile documents should be treated distinctively from those in tweets. (iii) We reveal that to adjust amount of information from social neighbors affects estimation accuracy. It is known that “birds of a feather flock together” relationships exist in social media [5, 19], however it is not clear how much information about the neighbors should be used together with those of the target user. We experiment three adjustment levels and show that our method, which utilizes the target user’s profile document and tweets and the neighbors’ profile documents (not including tweets), achieved the best accuracy.

Outline. Section 2 describes related work on estimating user attributes. Section 3 presents the results of our analysis of Twitter data and reveals the necessity of user attribute estimation. Section 4 details our proposed methods. Section 5 presents the results of our experiments. Section 6 summarizes the contributions of this study.

2 Related Work

We present related work on the estimation of user attributes from a user’s contents and social graph. We then describe the scope of this study and the advances over existing methods.

2.1 Estimation of User Attributes from the Contents

This section details recent studies that focused on contents, i.e. a profile document and tweets. Ikeda et al. [8] extracted bag-of-words features using the Akaike Information Criterion (AIC) and estimated user attributes by Support Vector Machines (SVM). Their experiments tackled three attributes (gender, age, and location), and the results showed 88 % accuracy with regard to gender. Ikeda et al. method is used as a baseline in Sect. 5.1. The differences between their method and our method are a labeling method (manual or automatic) and an estimator (SVM or logistic regression), but an estimator is aligned with logistic regression in the experiment of Sect. 5.1 to compare the difference of a labeling method. Rao et al. [15] used n-grams or sociolinguistic features and estimated four user attributes (gender, age, location, and political orientation) by SVM. Their proposed method achieved 70–80 % accuracy in estimating

the attributes, and they reported that Twitter-specific features (number of followers, number of friends,⁶ friends/followers ratio, reply ratio, number of tweets, and number of retweets⁷) are not useful. Cheng et al. [3] proposed two methods for estimating the city-level location of users: a probability model that is based on the correlation between a location and each word in the tweets and a grid-based neighborhood smoothing model for adjusting the estimation of the user location. Their method can estimate the location of 51 % users with an error range of 100 miles, from 100 tweets. Eisenstein et al. [6] estimated the location of users based on an idea similar to that of Cheng et al. [3] that there exists a strong correlation between a word and a particular region. It is different from Cheng et al. work in that it uses a generative model that estimates the latent topics and regions together. Burger et al. [2] estimated the gender of users by using a supervised learning method that employs the both word and character n-grams as features. They achieved 92 % accuracy by using the feature set of tweets, profile documents, screen name, and name. Their method performed better than manual estimation based on the Amazon Mechanical Turk⁸ in terms of accuracy. Pennacchiotti et al. [14] estimated three attributes (the political orientation, race, and affinity for Starbucks Coffee) of users. They use the profile documents, tweeting tendency, and characteristic words in tweets as features. They update attribute-class label information by using the social graph and estimate user attributes by the Gradient Boosted Decision Trees (GDBT). Chu et al. [4] used the tweeting tendency, the contents of tweets, and the profile as features to distinguish human from bots by Linear Discriminant Analysis. Mislove et al. [13] compared Twitter user distribution with the actual population distribution with regard to gender, race, and location attributes. They showed that there is a bias in the distribution of users on Twitter.

2.2 Estimation of User Attributes from the Social Graph

This section describes recent studies on estimating user attributes by using the attributes of neighbors on social graph. Zamal et al. [19] tried various methods that combine a user's feature and the averaged value of the same feature for the user's neighbors. Zamal et al. method is used as a baseline in Sect. 5.4. Their method is quite different with our method in the respects of a labeling method, an estimator, and features. But the respect of how to use neighbors' information is the same. We compared the method which uses neighbors' tweets (Zamal et al. method) with our methods in Sect. 5.4. Mislove et al. [12] used the social graph of Facebook to estimate the user attributes such as the enrollment year and the department. They set the nodes that have the same attribute value as seed nodes and add the remaining nodes to these so as to increase a modularity-based value. Wen et al. [17, 18] estimated interests from large-scale monitored data, such as emails, instant messages,

⁶ A friend is another Twitter user whom you are following.

⁷ A Tweet by another user, forwarded to you by someone you follow.

⁸ <https://www.mturk.com/>.

social bookmarks, and file-sharing data. They proposed a propagation model that is weighted by the number of communication events. He et al. [7] created a small group, called *homogeneous societies*, that reflects reality and used the Bayesian Network approach to estimate user attributes of blog. Zheleva et al. [20] tested whether user attributes can be estimated using friends’ and group’s information. They reported that group’s information achieved higher accuracy than friends’ information. From the viewpoint of privacy protection, Lindamood et al. [10] examined the impact of hiding a part of user attributes or user’s friends information so as to prevent user attributes from being estimated by others.

2.3 Scope of This Study

The main purpose of this study is to construct a less human effort, low computational cost, but highly accurate estimator for practical use. For less human effort, we focus on labeling methods that do not require manual labeling of training data. For computational cost savings, we only utilize bag-of-words features, one-hop neighbors of the target user, and logistic regression models. For high accuracy, we add features of profile documents and social neighbors to existing tweets’ feature. Therefore, our purpose is different from those of other studies that use manual labeling, Twitter-specific features, or propagate user attributes by using the entire social graph.

3 Analysis of the Twitter User’s Profile

Twitter users can write self-introduction sentences (a profile document) in free-form text. It is not necessary to estimate the user attributes if these are specified in the profile document. Therefore, we analyzed Twitter data to determine how many users entered attributes in the profile document.

3.1 Structures of the Twitter Profile Field that We Used

Table 1 shows the structures of the Twitter profile field used in our study. *Description* and *location* are fields that are likely to hold user attributes. *Statuses_count* indicates

Table 1 Structures of the Twitter profile field that we used

Field name	Explanation
<i>description</i>	Field to hold self-introduction sentences
<i>location</i>	Field to hold location
<i>statuses_count</i>	Field to show total number of tweets
<i>url</i>	Field to hold website

the total number of tweets of a user. *Url* is a field for an external URL such as blog URL or the user’s website’s URL.

3.2 Analytical Methods and Results

Table 2 details the data analyzed. We collected over 4.6 million Japanese Twitter users who have posted at least one tweet during March 2012. We analyzed how many users filled in the *description* and the *location* fields and whether the entered text contained words that indicate user attributes. Attribute words expected were manually predefined and regular expressions were used for matching. Additionally, we calculated the description rate for each attribute.

Table 3 shows the results of the analysis. Over 82 % of users entered text in the *description*. Unfortunately, the usage of specific words was low; the description rate for age was only 3.34 %. This description rate is overly optimistic because of the presence of extraction noise. For example, age of the user’s children or pets tend to be described in the user’s profile document and this data must be seen as noise. Although the *location* is a special field to describe the user’s location attribute and its description rate is relatively high, extraction noise is present, e.g. multiple descriptions of location that a user lived in the past.

From this analysis, we found that most users write something in their profile document but rarely their attributes. This makes it essential to estimate the user attributes. We also found that manual checking of the extraction results is inevitable because applying text matching method to a free-form text tends to yield a lot of noise.

We then counted the number of *statuses_count* and analyzed how many tweets were, on average, available as an estimation resource. The mean value was 68.1 and the median value was 553. A tweet is shorter than a blog article because it is

Table 2 The data used to analyze Twitter user’s profile

Field name	Value
API level	Gardenhose (10 % random sampling)
Data collection period	3/1/2012–3/31/2012
No. of unique Japanese users	4,638,441

Table 3 Description rate of user attributes

Field name	No. of described users	Description rate (%)
Any description	3,827,885	82.53
Gender	353,558	7.62
Age	154,900	3.34
Occupation	631,626	13.62
Location	1,158,570	24.98

limited to 140 characters, and the mean value of *statuses_count* is low. Therefore, the estimation accuracy may decrease if the target user does not have enough number of tweets.

4 Proposed Methods

In the following sections, we propose a labeling method, content-based methods, and social-graph-based methods to solve the problems described in Sect. 3.

4.1 Labeling Method

It is difficult and time-consuming to label training data manually. Although we can use regular expressions to label automatically, this method contains extraction noise as shown in Sect. 3, so manual confirmation is needed. The reason why labeling fails so often is that profile documents hold free-form text. On the other hand, blog’s profile fields are usually provided for each attribute and multiple choice selections are provided for user entry. It is easy to extract user attributes from these fields by using a rule-based method. Therefore, automatic learning is available by using the users who have both blog and Twitter accounts, we call them TwiBlo users, as training data.

We counted up the number of distinct URLs in the *url* field for each domain from the users shown in Table 2. Top 10 results are showed in Table 4. To see the results, the 1st rank domain has about 160 thousand TwiBlo users, and the top 10 blog sites have over 240 thousand TwiBlo users. This is 10–100 times higher than the amount of manually labeled data [8, 14, 15]. Manual labeling is difficult and time-consuming, so the amount of data tends to be less than several thousand. In general, estimation accuracy increases with the amount of training data, so this labeling method is expected to achieve higher accuracy than manual labeling methods.

Here, we explain the flow of this labeling method. It first finds users who have both Twitter and blog accounts (TwiBlo users) by using the URL described in the *url* field. It then extracts the attributes that a user specified in his/her blog as true labels of the training data in Twitter about the user. The extraction rules are predefined for each blog site and extraction script outputs user attributes when it is input the URL described in the *url* field.

Table 4 Domain names present in the url field

Rank	Domain name	No. of users	Rank	Domain name	No. of users
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

Burger et al. [2] also used TwiBlo users as training data, but they verified the effectiveness of this method only in gender and conducted validity check manually by using one thousand randomly sampled users. By contrast, we demonstrate the effectiveness of this method with respect to four attributes and conduct verification experiments in Sects. 5.1 and 5.2.

4.2 Content-Based Methods

We lack enough information to estimate user attributes depending on the target user's tweeting history when only his/her tweets are used as training data. Accurate estimation requires more information than is contained within the tweets, so we consider the use of profile documents together with tweets. There is only one profile document per user, but it is a high quality source of information since it may hold actual user attributes, so estimation accuracy can be increased by using it together with tweets as the training data. However, we do not know what is the best way to mix/combine profile documents and tweets. Therefore, we create following nine methods and determine which is best in Sect. 5.

MIX. This estimator counts words from profile documents and tweets without distinction; this is equal to consider a profile document as a tweet.

JOIN. This estimator treats words from profile documents differently from those in tweets. Thus, the number of feature dimension is greater than the estimator using only tweets.

AVG. Creates two estimators, one each from profile documents and tweets. The outputs of the estimators are averaged.

MAX. Creates two estimators, one each from profile documents and tweets. The output with maximum value is adopted.

VAR. Creates two estimators, one each from profile documents and tweets. The output with maximum variance is adopted.

DEF. Creates two estimators, one each from profile documents and tweets. The ratio of margin of defeat between 1st and 2nd classes is calculated for each estimator's output and the output with the minimum ratio is adopted.

KIND. Creates two estimators, one each from profile documents and tweets. The equations are as follows:

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

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

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

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

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

Let u be a user with the user's profile document u_p and tweets u_t . As indicated by Eq. (1), final estimation probability P is obtained by aggregation of estimation probabilities from the profile document P_p and tweets P_t weighted by reliability scores R_p and R_t , respectively. R_p and R_t are calculated by self-information I_p and I_t that are obtained by the log ratio of the kind function's value and the total number of kinds of features $|F_p|$ and $|F_t|$ shown as Eqs. (2) and (3), respectively. F_p and F_t are extracted in a feature selection step; in this study, we employ the Akaike Information Criterion (AIC) [1] as the feature selection method [8, 11] for all the content-based methods. The kind function returns the number of kinds of features used to estimate the target user's attributes. α is a constant value to prevent the logarithm from being zero; we set it to 1.

AIC. Creates two estimators, one each from profile documents and tweets. Equations (2) and (3) in **KIND** are replaced with (4) and (5), respectively:

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

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

where `set` is a function that returns the set of features contained in the input text, and `aic` is another function that returns the AIC value of the input feature f , which is calculated in the feature selection step.

RANK. Creates two estimators, one each from profile documents and tweets. Equations (2) and (3) in **KIND** are replaced with (6) and (7), respectively:

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

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

$$\text{rank}(f) = \frac{|F|}{\text{index}(f)}, \quad (8)$$

where `rank` is a function that returns the rank value of the input feature $f \in F$ (F would be F_p or F_t) shown as Eq. (8). The rank value is obtained by the ratio of the index function's value and the total number of kinds of features. The index function returns the index of features sorted by AIC values, which are calculated in the feature selection step, in descending order; initial value is 1.

4.3 Social-Graph-Based Methods

Neighborhood users on social graph are known to tend to have similar attributes; this is called the “birds of a feather flock together” effect [5]. Utilizing the information about neighbors can make user attribute estimation more stable even if the user does not have enough information to permit his/her user attributes to be estimated.

We propose new methods that use the information of the target user’s neighbors by extending Zamal et al. methods [19]. The main difference of our methods is the information used from the target user and the social neighbors. We set three information levels as follows.

PR. Uses profile documents only (PProfile).

TW. Uses tweets only (TWweets).

TP. Uses both profile documents and tweets (Tweets and Profile).

We tried various combinations of these information levels between a target user and his/her social neighbors. Here, Zamal et al. method is equivalent to the case of applying TW to both the target user and social neighbors. However, it is not exactly the same in terms of features used; they used various features such as n-grams or tweeting tendency, but we use only bag-of-words features.

The methods of selecting social neighbors and combining those features with the target user’s features are the same as in Zamal et al. as follows.

ALL. All social neighbors are used.

MOST. N -most popular social neighbors are selected. The popularity is assessed in terms of the number of followers of the user. We set 10 for N in our study (following N is the same value).

LEAST. N -least popular social neighbors are selected. It is based on an expectation that a user dares to follow unpopular users because they are real friends.

CLOSEST. N -most closest social neighbors are selected. Closeness is determined by the number of conversations (mentions) from a target user to the neighbors.

The features of the selected neighbors are averaged and then combined with the target user’s feature. AVG and JOIN are employed as the combination methods. AVG averages the feature sets of the target user and the social neighbors, and JOIN treats those sets distinctively and joins them.

There are various ways to construct a social graph, e.g. using the relation of friends, followers, and mentions. We use mention relations because of its characteristics and Twitter API’s restrictions. It is a closer relationship than friends/followers relationships because mention is an active action. Twitter provides REST API which returns the friends/followers of a user, but its rate limit is 350 calls per hour now, and will fall to 15 calls per 15 min from May 2013. It is hard to get friends/follower networks given this API restriction. On the other hand, mentioning relationship is extracted from tweets collected by the streaming API, so it is not necessary to use the REST API.

5 Evaluation Experiments

We conducted experiments to evaluate the effectiveness of our proposed methods. Table 5 details the data used in the evaluation experiments, and Table 6 shows classes in each attribute. We crawled about 86 thousand TwiBlo users, who are Twitter users associated with a blog account, and up to 200 tweets, excluding RTs, per user. Blog users and their articles are different from TwiBlo user’s data and this blog data are used in Sect. 5.2 for evaluating our labeling method. We used AIC for the feature selection method and LIBLINEAR⁹ with L2-regularized logistic regression (primal) for constructing the estimators. The reason we used logistic regression is to get probabilistic output. Based on the results of a preliminary experiment, we set the number of features and the cost parameter, which is C parameter for L2-regularized logistic regression (primal). Five thousand and thirty thousand features were extracted from profile documents and tweets, respectively. The cost parameter was set to 1.

Table 5 Experimental data

	Field name	TwiBlo users ^a	Blog users
	Gender	71,129	49,739
	Age	36,234	17,689
No. of users	Occupation	41,920	37,427
	Interests	20,846	7,417
	Total	86,183	65,873
No. of blog articles		796,583	626,903
No. of tweets		15,124,094	–

^a Twitter users associated with a blog account

Table 6 Classes in each attribute

Attribute	Classes (number of class)
Gender	Male and female (2 classes)
Age	10s, 20s, 30s, and over 40s (4 classes)
Occupation	Company employee, government employee, self-employed, job-hopping part-time worker, housewife, graduate or undergraduate student, middle or high-school student, and others (8 classes)
Interests	Reading, cooking, traveling, gourmets, sweets, computers, internets, games, musics, sports, pets, comics and animations, television and movies, politics and economics, learning and education, health maintenance, career development, finance, entertainment world, and fashion (20 classes)

⁹ <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>.

5.1 Comparison Against Manual Labeling

We conducted an experiment to compare the estimation accuracy yielded by our labeling methods and three other labeling methods. REGEXP is a labeling method that matches specific words which associate user attributes to words contained in profile documents by using regular expressions. HUMAN is another labeling method that first author surveys REGEXP’s output and retains only correct results; it corresponds to Ikeda et al. method [8]. D1000 is the same as our labeling method (DIRECT), but its data size is equal to REGEXP and HUMAN.

We targeted gender and age attributes and based the evaluation on special 5-fold cross-validation. Figure 1 shows an overview of the experiment. TwiBlo users’ data were divided into five blocks, one was used as test data for all four methods, so all methods evaluated the same data. On the other hand, the training data were different for each of the four methods. DIRECT used the rest of the data to construct an estimator (model) when TwiBlo users’ data were folded, and D1000 used the same data but its size was restricted to 1,000 for each class by random sampling. REGEXP and HUMAN used the data obtained by each method and their size was 1,000 for each class. The same operation was repeated for the number of splits.

The results are shown in Table 7. The estimation accuracy of both gender and age attributes rises gradually in the order of REGEXP, HUMAN, D1000, and DIRECT. Though the amount of data was the same, D1000 was more accurate than HUMAN. It is because HUMAN used minority and special users who specified their attributes

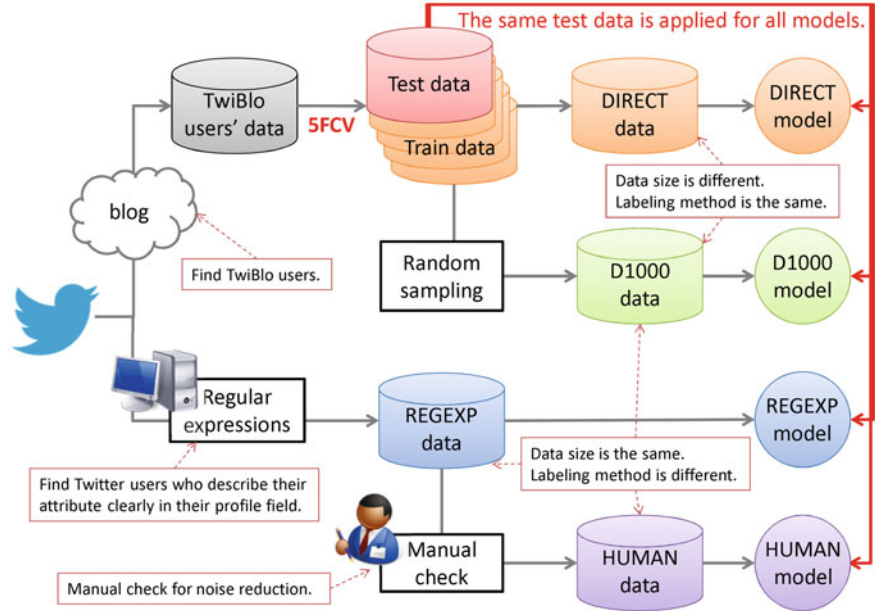


Fig. 1 An overview of the experiment in Sect. 5.1

Table 7 Accuracy with each labeling method

	REGEXP	HUMAN	D1000	DIRECT
Gender	72.59	82.32	89.39	94.50
Age	59.49	61.86	67.72	76.28

in their profile document. DIRECT achieved the best accuracy because its training data size was 35 and 9 times larger than the others in regard to gender and age, respectively. While it is difficult and time-consuming to collect training data manually, our labeling method can collect large amounts of training data automatically. Therefore, our labeling method can automatically make a more accurate estimator than the manual-based method.

5.2 Validation of Using Labels from Blog

Our labeling method learns from user tweets and the true labels found in his/her blog’s profile. When a user lies about his/her attributes in blog or writes completely different content between Twitter and blog, learning does not go well. For example, a user specifies travel as his interest attribute in his blog’s profile, but uses Twitter as a communication tool and his tweets contain no mention of travel.

We tried to remove from the training data users who entered different content in Twitter and blog. We tested two types of filtering method. The first used an estimator learned from randomly collected blog articles. It estimates user attributes from his/her tweets and blog articles, and filters out users whose estimation results differ from the blog’s profile. BOTH is the case that all three factors (estimation results from his/her tweets and blog articles, and his/her blog’s profile) are the same. BLOG and TWIT are the case that the estimation results from blog articles and tweets are the same as the blog profile and the other factors are different from the blog profile, respectively. These filtering methods depend on the accuracy of the estimator, so the second filtering method uses cosine similarity (COS) rather than an estimator. Bag-of-words word frequency vectors were created from his/her tweets and blog articles, and users whose cosine similarity between both vectors was less than 0.8 were dropped. We also evaluated the transfer learning method (TRANS) and the non-filtering method (DIRECT) which equals our labeling method. TRANS uses blog articles and tweets as training data and test data, respectively.

We conducted experiments based on special 5-fold cross-validation. Figure 2 shows an overview of the experiment. We divided TwiBlo users’ data into five blocks, one was used as the test data, and the remaining data were filtered by the methods described above; the filtered results were used as training data. DIRECT is not filtered and TRANS’s training data are different from the other methods, but all methods had the same test data. The same operation was repeated for the number of splits.

The results are shown in Table 8. Accuracy is shown as percentage and the amount of training data after filtering is given by the bracketed number. DIRECT achieved the best accuracy in all attributes except for interests. BLOG and COS achieved the

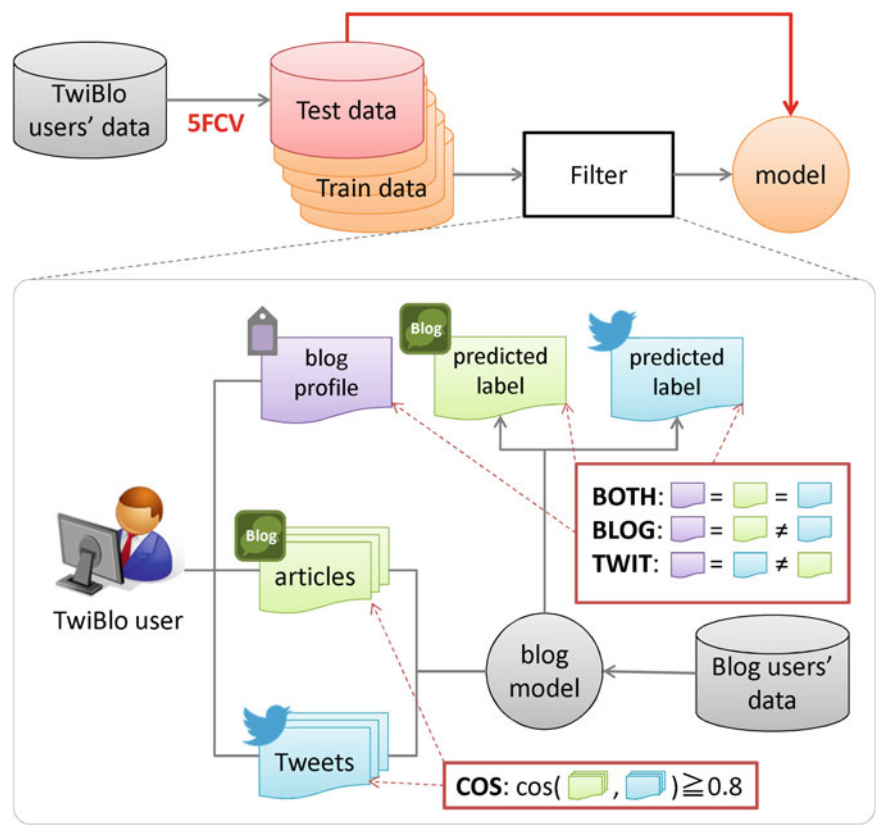


Fig. 2 An overview of the experiment in Sect. 5.2

Table 8 Accuracy with each filtering method

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

best accuracy in interests attribute, but the difference with DIRECT, 0.03 %, was not statistically significant in McNemar’s test. COS creates a small training data set, but its accuracy does not decrease commensurately. TRANS failed to achieve good accuracy because test and training data had different domains.

From these results, we conclude that falsification of blog profile entries and the difference in content between Twitter and blog is little. Therefore, we are able to apply our labeling method without filtering the training data, and this achieves high accuracy for all attributes examined here.

5.3 Comparison of Methods that Use Profile Documents

To confirm that profile documents must be used together with tweets to improve the estimation accuracy, we conducted experiments on the various methods shown in Sect. 4.2. In addition to these methods, we also evaluated the method that uses only profile documents (PROF) and the method that uses tweets only (TWEET). TwiBlo users’ data were used for the experiments, and four attributes (gender, age, occupation, and interests) were evaluated with 5-fold cross-validation.

The results are shown in Table 9 and Figs. 3, 4, 5, and 6. The table lists estimation accuracies for each method and attribute, and the average of the ranks for each method. Figures plots accuracy–coverage curves (ACC); x-axis is coverage and y-axis is accuracy. The best possible prediction method would yield points in the upper

Table 9 Accuracy comparison of methods that use profile documents

	PROF	TWEET	MIX	JOIN	AVG	MAX	VAR	DEF	KIND	AIC	RANK
Gender	78.98	94.20	94.20	94.46	94.03	94.03	94.03	94.03	94.46	94.53	94.60
Age	60.43	72.26	72.90	73.91	73.20	72.95	72.89	72.63	73.43	73.45	73.36
Occupation	52.29	58.71	58.84	61.30	61.81	61.13	60.74	61.21	61.49	61.45	61.46
Interests	54.00	56.92	57.73	61.56	60.51	59.88	59.55	60.23	60.29	60.38	60.18
AOR ^a	11	9	7.5	2.75	4.125	7.125	8.125	7.125	3	2.75	3.5

^a Average of the ranks

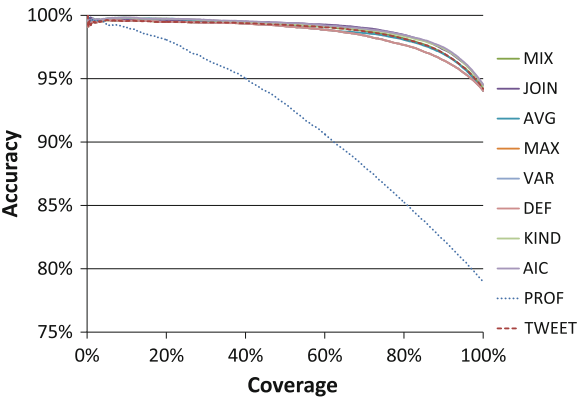


Fig. 3 Accuracy–coverage curve of gender

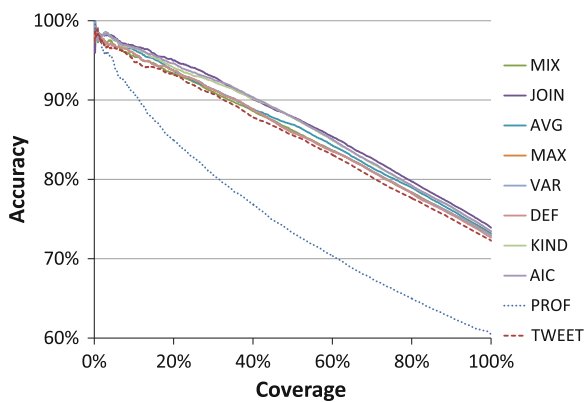


Fig. 4 Accuracy-coverage curve of age

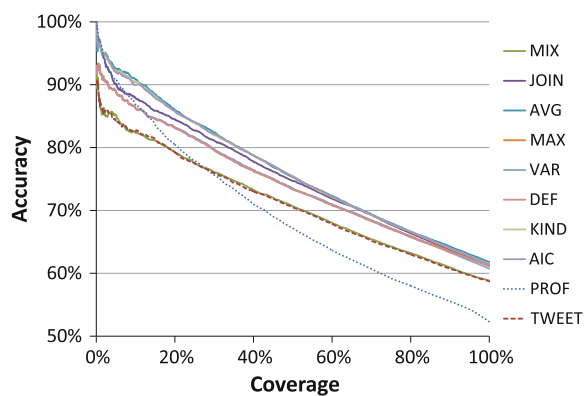


Fig. 5 Accuracy-coverage curve of occupation

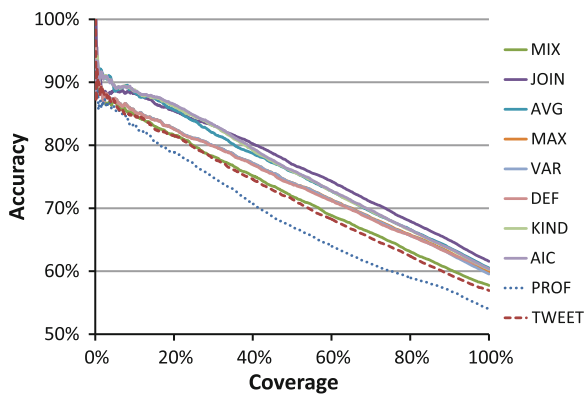


Fig. 6 Accuracy-coverage curve of interests

side of ACC space; it means always 100 % accuracy in any coverage. ACC is good to be plotted near the top right corner.

JOIN and AIC achieved the best average of the ranks as shown in Table 9. JOIN achieved the best accuracy in gender and interests attributes, and AIC offered high and stable accuracy but not the highest accuracy. With both methods, a significant difference was observed in McNemar's test at a significance level of 5 % compared to TWEET in all attributes. Therefore, we propose using JOIN or AIC method, either of which use profile documents together with tweets, to estimate user attributes. JOIN achieved higher accuracy than MIX, so words in profile documents should be treated differently from words in tweets.

The table shows the accuracy only at the point of 100 % coverage, but the figures allow us to grasp more details. The figures show that the accuracy tends to decrease monotonically, and the accuracy ranking of methods remains basically unchanged as coverage changes. The difference in the accuracies of the methods appears most strongly in the attributes that have many classes, i.e. occupation (8 classes) and interests (20 classes). PROF has higher accuracy than TWEET at low coverage levels as shown in Fig. 5. This indicates that the number of users who specify helpful words to estimate occupational attribute in their profile document is larger than the other attributes. This result presents that the effect of using profile documents varies according to attributes.

5.4 Evaluation of Social-Graph-Based Method

We conducted experiments on the methods shown in Sect. 4.3, and the results are shown in Table 10. The methods tested are written as [the combination method of feature sets (AVG or JOIN)]-[the target user's information type (TW or TP)][the neighbors' information type (PR, TW, or TP)]-[the selection method of the neighbors (ALL, MOST, LEAST, or CLOSEST)], e.g. AVG-TWTW-ALL. We also evaluated the methods that use the target user's profile document only (PROF), tweets only (TWEET), both the profile document and tweets (TWEPRO; it is the same as JOIN in Sect. 4.2), and the neighbors' tweets only (NBR).

We based experiments on 10-fold cross-validation, and the table shows accuracies for each method and user attribute. Accuracy values higher than TWEPRO, which is a baseline method that does not use neighbors' information, are bold faced to simplify comparison with the content-based method, and the highest accuracy for each attribute is indicated by an asterisk. The experimental data are the same as shown in Table 5, but size was confined to 200 for each class due to the difficulty of collecting large numbers of neighbors' tweets. So, accuracies are low compared to Table 9 because of the reduction of training data.

Table 10 indicates that almost social-graph-based methods yielded higher accuracy than TWEPRO (content-based method) for age attribute, and NBR yielded lower accuracy than TWEPRO for all attributes except age. These indicate that the

Table 10 Accuracy comparison of each social-graph-based method

	Gender	Age	Occupation	Interests
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-TTPR-ALL	86.75	65.00	48.09	54.45*
JOIN-TTPR-MOST	86.75	63.88	48.21	54.27
JOIN-TTPR-LEAST	86.25	64.63	48.41	53.64
JOIN-TTPR-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

strength of homophily¹⁰ depends on the attributes and that of age is strong. JOIN achieved higher (and stably) accuracy than AVG, so words from the target user and the neighbors should be treated differently. JOIN-TTPR-* is superior in accuracy to TWEPRO and Zamal et al. methods (*-TWTW-*), and achieved the best accuracy in gender and interests attributes. In general, the estimation accuracy increases with increasing the amount of training data, but the neighbors data can be a noise when the estimating attribute has weak homophily. These results showed that the estimation accuracy becomes stable for all attributes by using only information from the profile documents of the neighbors. Unfortunately, we were not able to find a clear winner among the neighbors' selection methods.

¹⁰ The tendency of individuals to associate and bond with similar others.

6 Conclusion

We propose new methods 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. Although there are many studies on estimating user attributes, we proposed new methods and showed new knowledge as indicated by the following three points. First, we investigate a labeling method that finds the users associated with a blog account (TwiBlo users) and uses their profile attributes on blog as true labels of training tweet data. We confirm that using the blog labels achieved higher accuracy than manual labeling and pattern matching methods, with respect to four attributes (gender, age, occupation, and interests). Additionally, our labeling method does not require any filtering of training data, because our experiments showed that the influence of blog's profile falsification and the difference in the Twitter and blog contents are slight. Second, we validate the best way to combine bag-of-words features of profile documents and tweets. We evaluate nine combining methods and show that words in profile documents should be treated distinctively from those in tweets. Our experimental results revealed that JOIN and AIC, described in Sect. 4.2, are the best choices. Third, we reveal that to adjust amount of information from social neighbors affects estimation accuracy. We experiment three adjustment levels and show that our method, which uses the target user's profile document and tweets and the neighbors' profile documents, experimented as JOIN-TPPR-* in Sect. 5.4, achieved the best accuracy.

We proposed a labeling method using blog as an example of a true label propagation source, however other media such as Facebook can be used instead of blog if the media holds user attributes as true labels. In future work, we will investigate the impact on estimation accuracy of using other media.

References

1. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19(6):716–723
2. Burger JD, Henderson J, Kim G, Zarrella G (2011) Discriminating gender on Twitter. In: *EMNLP*, pp 1301–1309
3. Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating Twitter users. In: *CIKM*, pp 759–768
4. Chu Z, Gianvecchio S, Wang H, Jajodia S (2010) Who is tweeting on Twitter: human, bot, or cyborg? In: *ACSAC*, pp 21–30
5. Conover M, Gonçalves B, Ratkiewicz J, Flammini A, Menczer F (2011) Predicting the political alignment of Twitter users. In: *SocialCom*, pp 192–199
6. Eisenstein J, O'Connor B, Smith NA, Xing EP (2010) A latent variable model for geographic lexical variation. In: *EMNLP*, pp 1277–1287
7. He J, Chu WW, Liu ZV (2006) Inferring privacy information from social networks. In: *ISI*, pp 154–165
8. Ikeda K, Hattori G, Matsumoto K, Ono C, Higashino T (2012) Demographic estimation of twitter users for marketing analysis. *IPSJ Trans Consum Devices Syst* 2(1):82–93

9. Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter power: tweets as electronic word of mouth. *J Am Soc Inf Sci Technol* 60(11):2169–2188
10. Lindamood J, Heatherly R, Kantarcioglu M, Thuraisingham B (2009) Inferring private information using social network data. In: *WWW*, pp 1145–1146
11. Matsumoto K, Hashimoto K (1999) Schema design for causal law mining from incomplete database. In: *DS*, pp 92–102
12. Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) You are who you know: inferring user profiles in online social networks. In: *WSDM*, pp 251–260
13. Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist JN (2011) Understanding the demographics of Twitter users. In: *ICWSM*
14. Pennacchiotti M, Popescu AM (2011) Democrats, republicans and starbucks aficionados: user classification in Twitter. In: *KDD*, pp 430–438
15. Rao D, Yarowsky D, Shreevats A, Gupta M (2010) Classifying latent user attributes in Twitter. In: *SMUC*, pp 37–44
16. Trusov M, Bucklin RE, Pauwels K (2009) Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *J Mark* 73(5):90–102
17. Wen Z, Lin CY (2010) On the quality of inferring interests from social neighbors. In: *KDD*, pp 373–382
18. Wen Z, Lin CY (2011) Improving user interest inference from social neighbors. In: *CIKM*, pp 1001–1006
19. Zamal FA, Liu W, Ruths D (2012) Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In: *ICWSM*
20. Zheleva E, Getoor L (2009) 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

Online Social Media Analysis and Visualization

Kawash, J. (Ed.)

2014, XVI, 233 p. 94 illus., 76 illus. in color., Hardcover

ISBN: 978-3-319-13589-2