Knowledge-Based Systems 51 (2013) 35–47

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Twitter user profiling based on text and community mining for market analysis

Kazushi Ikeda [a,*], Gen Hattori [a], Chihiro Ono [a], Hideki Asoh [b], Teruo Higashino [c]

[a] KDDI R&D Laboratories, Inc., 2-1-15, Ohara, Fujimino, Saitama 356-8502, Japan
[b] National Institute of Advanced Industrial Science and Technology, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
[c] Graduate School of Information Science and Technology, Osaka University Yamadaoka, 1-5, Suita-shi, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Received 22 August 2012
Received in revised form 27 June 2013
Accepted 29 June 2013
Available online 12 July 2013

Keywords:
Web mining
Market analysis
User profiling
Twitter
Text analysis
Community analysis
Machine learning

ABSTRACT

This paper proposes demographic estimation algorithms for profiling Twitter users, based on their tweets and community relationships. Many people post their opinions via social media services such as Twitter. This huge volume of opinions, expressed in real time, has great appeal as a novel marketing application. When automatically extracting these opinions, it is desirable to be able to discriminate discrimination based on user demographics, because the ratio of positive and negative opinions differs depending on demographics such as age, gender, and residence area, all of which are essential for market analysis. In this paper, we propose a hybrid text-based and community-based method for the demographic estimation of Twitter users, where these demographics are estimated by tracking the tweet history and clustering of followers/followees. Our experimental results from 100,000 Twitter users show that the proposed hybrid method improves the accuracy of the text-based method. The proposed method is applicable to various user demographics and is suitable even for users who only tweet infrequently.

1. Introduction

Recently, due to the widespread popularity of the Internet, many people state their opinions via social media services. In particular, Twitter [1] is a suitable platform for real-time, casual communication. Many Twitter users post opinions about products, services, and TV programs. It is essential for companies to make efforts to improve their products and services based on their customers' requirements. As a means of using user opinions for marketing, reputation analysis technologies have recently attracted a great deal of attention [2,3]. Compared to previous marketing approaches based on questionnaire surveys, online opinion analysis has many advantages, including real-time feedback, low cost, and high volume. User demographics such as age, gender, and residence area are also essential for marketing analysis, since opinions vary with user demographics. For example, functions of mobile phones that are popular among young people are often found awkward to use by the elderly. Since most Twitter users do not state their demographic information, it has been impossible to extract opinions for individual user demographic segments (such as teens, twenties, or thirties). Several text-based approaches have been proposed to extract user demographic information [4–6]. However, only few proposals for large-scale and practical marketing analysis applications to perform demographic estimation exist due to difficulties in improving the effectiveness and accuracy to a level sufficient for practical use. Considering practical use, we realized that a general approach is required for estimating wide varieties of demographics such as age, gender, area and other categories. An estimation method targeting users with few tweets such as followers of corporate accounts is also important.

To solve these problems, we propose a hybrid of a text-based method and a community-based method for the demographic estimation of Twitter users. The text-based method estimates the demographics of users whose tweets contain sufficient text features. For all other users, the community-based method analyzes the followers/followees whose tweets contain plentiful text features. The hybrid method covers almost all users by making the most of the Twitter platform, including both tweets as text information and followers/followees as community information. In the text-based method, characteristic terms used by each demographic segment are automatically detected based on linguistic and statistical analysis by tracking the content of users' tweet histories. For example, users whose tweets often include terms such as "school," "classroom," and "examination" are presumed to be teens and students. In the community-based method, demographic

* Corresponding author.
E-mail addresses: kz-ikeda@kddilabs.jp (K. Ikeda), gen@kddilabs.jp (G. Hattori), ono@kddilabs.jp (C. Ono), h.asoh@aist.go.jp (H. Asoh), higashino@ist.osaka-u.ac.jp (T. Higashino).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.06.020

information is estimated from the follower/followee relations of the target user. In the proposed method, characteristic biases in the demographic segments of users are detected from the community groups constructed by clustering their followers and followees. A user can have several community groups, such as local friends, co-workers and hobby groups, where the members of each group have something in common such as age, gender and regional area.

Social opinions and demographic information are extremely attractive to businesses. For instance, product planners need to understand user requirements, customer support and service management departments need to monitor customer responses, advertising agencies want to deliver persuasive advertisements to target audiences, and broadcast TV directors need real-time feedback from the audience. In this paper, we focus on Japanese Twitter users. However, the algorithms of the proposed text-based and community-based methods are applicable to any language.

The rest of the paper is structured as follows. We outline related work in Section 2. We describe the proposed text-based method, community-based method and hybrid method for demographic estimation in Section 3 and the results of performance evaluations in Section 4, respectively. We conclude this paper in Section 5.

2. Related work

Extracting author information from the Web has been attempted for a long time. Table 1 summarizes the previous works. An extraction method for author information from Web sources is proposed for the purpose of judging whether the information is trustworthy [7]. Koppel et al. classify three author attribution problems [8]: (1) the profiling problem, where the challenge is to provide as much demographic or psychological information as possible about the author [4–6]; (2) the needle-in-a-haystack problem, where there are many thousands of candidates for each of whom we might have a very limited writing sample [9]; and (3) the verification problem, where the challenge is to determine whether the target is the author or not [10]. The problem that we tackle in this paper is related to (1).

Common approaches to author profile estimation from documents use the volume of each term in the document for classification. Argamon et al. estimate the authors' age, gender, native language, and personality from blogs and essays written by university students [4]. Estival et al. estimate age, gender, nationality, education level, and native language from English e-mails [5]. Pham et al. estimate age, gender, and area from Vietnamese blogs [6]. However, in these previous studies, the evaluations are only on small platforms, such as blogs, essays, or e-mails. With practicality in mind, we propose a profile estimation method on Twitter, which is one of the largest, most popular, and most internationally accepted platform among social media.

There are challenges associated with the author attribution problem of (2) and (3) on the Twitter platform [9,10]. Layton et al. show that the important threshold is 120 tweets per user, at which point adding more tweets per user gives a small but non-significant increase in accuracy for the author attribution problem [11]. Silva et al. show that markers include highly personal and idiosyncratic editing options, such as emoticons, interjections, and punctuation, which are often seen in casual SNS (Social Networking Services) such as Twitter [12]. In these studies, only text information is used, which is considered to limit accuracy. We propose a hybrid method comprising a text-based method and a community-based method, which enhances the accuracy.

Follower and followee relationships are regarded as directed links between two users. There has been some research reporting link-based document classification methods, such as for scientific papers based on co-citations [13] and for Web pages based on their hyperlinks [14]. Hybrid methods composed of text-based methods and link-based methods are reported to improve the accuracy of Web page classification [15–18]. Calado et al. show that classification methods based on co-citation are effective [15]. Qi and Davison show that the content and topic information of neighboring documents increases classification accuracy [16]. Zhang et al. propose an optimization method with text-based and graph structures for categorization of Web pages [17]. These contributions are helpful for improving the performance of text-based methods. The methods are designed on the assumption that a Web page belongs to only one category at a time. In the demographic estimation problem, however, a user has multiple demographics such as age, gender and area. In that case, existing clustering algorithms for Web pages are not simply applicable to the problem. Therefore, we propose a demographic estimation method targeting user communities.

We have previously proposed a text-based demographic estimation method for Twitter users and its application to broadcast TV programs [19]. In this previous work, the demographic estimation method is limited in terms of the minimum functions required for consumer usage. Only the basic demographic categories such as age, gender and area of residence are estimated. Estimation for users with few tweets is outside of its scope. Going beyond the previous work, we have conducted inquiry surveys in five depart-

Table 1
Summary of previous works related to extraction of author information.

Methods	Information source	Problems (target profiles)	Algorithms
Profile estimation			
Argamon et al. [4]	English blog and essay	Age, gender, native language and personality	Text-based classification
Estival et al. [5]	English e-mail	Age, gender, nationality, education level and native language	Text-based classification
Pham et al. [6]	Vietnams blog	Age, gender and area	Text-based classification
Ikeda et al. [19] (authors' previous work)	Twitter (Japanese tweet)	Age, gender and area	Text-based classification
Proposed method (this paper)	Twitter (Japanese tweet)	Age, gender, area, hobby, occupation and marital status	Hybrid of text-based and community-based
Other related work			
Kato et al. [7]	Web page	Increase credibility of information	Extract authors information
Abbasi and Chen [9]	Email, Web pages and chat	Find an author from thousands of candidates	Text-based information retrieval
Koppel et al. [10]	Literatures	Answer a given target text is or is not written by a given author.	Text-based classification
Layton et al. [11]	Twitter	Survey of the number of tweets required for the author attribution problems	Text-based classification
Silva et al. [12]	Twitter	Improve performance of author attribution problems	Text-based classification including markers

ments of four companies for disclosing the properties of the demographic estimation problems to design a practical market analysis application. As the result, we found out that among all demographic categories, age, gender, and area are most essential for all companies. In addition, more than three companies listed hobbies, occupation and marital status as important demographic categories. Therefore, in this paper, we target these demographic categories. We also found out that the estimation of users with few tweets such as followers of corporate accounts is also important, since a large number of users have few tweets or few feature can be extracted from their tweets. Fig. 1 shows the distribution of the number of tweet histories for 1000 randomly selected Twitter users using a streaming API. Less than half of them have 200 tweets or more. Text-based demographic estimation is not accurate enough for those users. To solve this problem, we propose an approach combining text-based and community-based methods to estimate multiple user demographics even for users with few tweets. We have evaluated the hybrid method for various demographic categories.

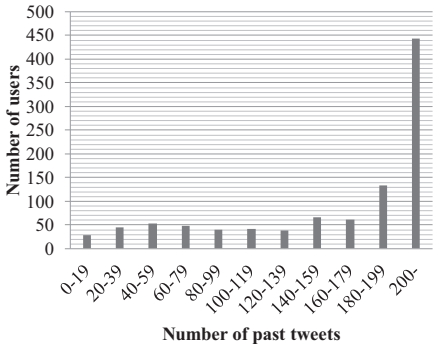

Fig. 1. Distribution of the number of Twitter users to the number of their tweet histories.

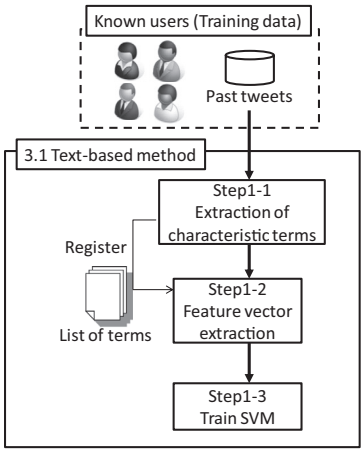

Fig. 2. Overview of the training phase of the proposed method.

3. Proposed method

The proposed hybrid method consists of a training phase and an estimation phase. Figs. 2 and 3 give an overview of the proposed method. In the training phase, the text-based method analyzes the tweet history of known users, extracts characteristic terms and trains SVMs with features of the used terms. In the estimation phase, the text-based method estimates the demographics of unknown users from their tweet history. The community-based method analyzes the follow/follower relations of unknown users, extract their communities and estimate the demographics of the communities. The hybrid method calculates the degree of membership for demographics of unknown users. In the following, we describe the details of the text-based method in Section 3.1, the details of the community-based method in Section 3.2 and the details of the hybrid method in Section 3.3.

3.1. Text-based method

The training phase described in Fig. 2 consists of the following three steps. Step 1–1: Extraction of the list of characteristic terms from the training data. Step 1–2: Extraction of feature vectors from the tweet history of known users. Step 1–3: Training the SVMs (Support Vector Machines). The estimation phase described in Fig. 3 consists of two steps. Step 1–4: Extraction of feature vectors from the tweet history of unknown users. Step 1–5: Estimation of the demographics of unknown users. This process can be regarded

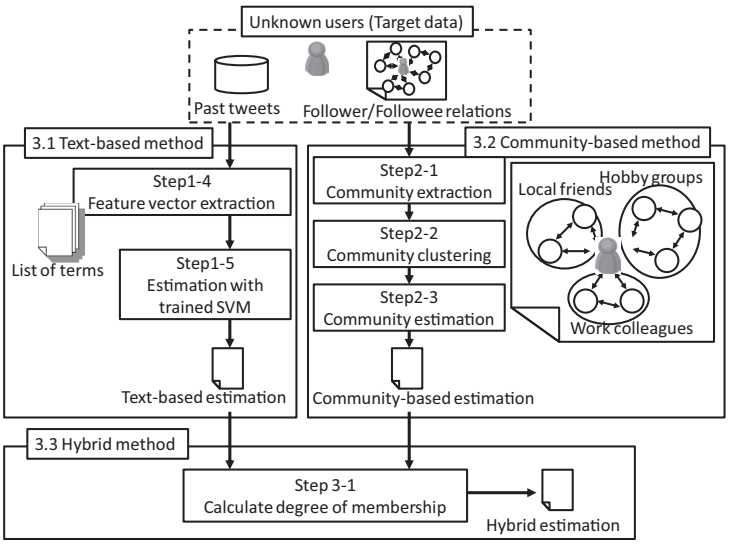

Fig. 3. Overview of the estimation phase of the proposed method.

as a kind of text classification, where the documents written by users are classified into any demographic segment.

On Twitter, users can introduce themselves in their user profile. A few users disclose their demographic information such as age, gender and area. However, most users do not disclose any demographic information. We surveyed the profiles of 244,345 Twitter users. The number of users with their profiles was 126,064 (51.6%). We randomly selected 100,000 users from those who disclosed some information in their profiles. The number of users who stated their age was 7338 (7.34%). Therefore, the ratio of users who disclose their age can be estimated as 51.6% × 7.34% = 3.8%. In the same manner, the ratio of users who stated their gender was 3.5%. We call users who state demographic information as known users, and users who do not state their demographic information as unknown users. We assume that there is no difference in the nature of tweets of known users and unknown users.

In Step 1–1, in order to realize a rapid demographic estimation for practical use, only selected characteristic terms are used. The number of varieties of the terms that appear in for example the latest 200 tweets of 1200 Twitter users is about 80,000. It requires much time to train an SVM with such a large feature vector. In order to reduce the time required for training the SVM, we narrow the term list based on the relevance of the characteristic term to the target demographic.

For evaluating the relevance, we use AIC (Akaike's Information criteria) [20]. AIC is a very popular criteria for statistical model evaluation, and by using AIC we can compare various models flexibly and universally. Here, we use AIC for comparing the dependent model (the appearance of the term t is related to the demographic) and the independent model (the appearance of the term t is independent to the demographic) [21]. The definition of AIC is stated in the following equation:

$$AIC = (-2)\log(\text{maximum likelihood}) + 2(\text{number of free parameters}) \quad (1)$$

The maximum likelihood of the dependent model ($MLL_{DM}(t)$) and of the independent model ($MLL_{IM}(t)$) can be computed from the counts of documents in the training dataset as:

$$\begin{aligned} MLL_{DM}(t) &= N_{11}(t)\log N_{11}(t) + N_{12}(t)\log N_{12}(t) \\ &\quad + N_{21}(t)\log N_{21}(t) + N_{22}(t)\log N_{22}(t) - N\log N \\ MLL_{IM}(t) &= N_p(t)\log N_p(t) + N(t)\log N(t) + N_n(t)\log N_n(t) \\ &\quad + N(\sim t)\log N(\sim t) - 2N\log N \end{aligned} \quad (2)$$

The definition of N_{11}, N_{12}, N_{21}, N_{22}, N_p, N_n, $N(t)$, $N(\sim t)$, and N is shown in Table 2. A document is a set of all tweets of a given user. Let D_p be a set of documents submitted by users within the target demographic segment (such as teens, male) and D_n be a set of documents submitted from other users. As shown in Table 2, N_{11} and N_{12} are the counts of documents in D_p and D_n, respectively where term t appears. N_{21} and N_{22} are the counts of documents in D_p and D_n, where term t does not appear. $N_n(t)$ and $N_p(t)$ are the counts of documents in D_p and D_n. $N(t)$ and $N(\sim t)$ are the counts of documents where term t appear and does not appear, respectively. N is the total number of the documents in the training datasets.

The number of free parameters of the dependent model is (number of columns) × (number of rows) − 1 = 2 × 2 − 1 = 3. The number of free parameters of the independent model is (number of columns − 1) + (number of rows − 1) = (2 − 1) + (2 − 1) = 2. Hence, the AIC for dependent model ($AIC_{DM}(t)$) and independent model ($AIC_{IM}(t)$) can be calculated with [21]:

$$\begin{aligned} AIC_{DM}(t) &= -2 \times MLL_{DM}(t) + 2 \times 3 \\ AIC_{IM}(t) &= -2 \times MLL_{IM}(t) + 2 \times 2 \end{aligned} \quad (3)$$

The relevance of the term t to the demographic segment can have two patterns, positive relevance and negative relevance. Positive relevance means that the term t appears frequently in tweets of users in the target segment. Negative relevance means that the term t appears frequently in tweets of users not in the target segment. In order to discriminate the two cases, we define the evaluation of relevance $E(t)$ as shown in Eq. (4). $E(t)$ returns positive values for positive relevance and negative values for negative relevance [22].

$$E(t) = \begin{cases} AIC_{IM}(t) - AIC_{DM}(t) + 2, & \text{if } \frac{N_{11}(t)}{N(t)} > \frac{N_{12}(t)}{N(\sim t)} \\ AIC_{DM}(t) - AIC_{IM}(t) - 2, & \text{if } \frac{N_{11}(t)}{N(t)} \leqslant \frac{N_{12}(t)}{N(\sim t)} \end{cases} \quad (4)$$

As an example, Table 3 shows the frequencies of appearance of the terms "examination," which is characteristic for teens, "beer," which is characteristic of non-teens, and "bucket," which uniformly appears in both teens and non-teens. "Examination" has a positive $E(t)$ value due to its positive relevance in teens' documents. "Beer" has a negative $E(t)$ value due to its negative relevance in teens' documents. "Bucket" has a low $E(t)$ value due to its uniform distribution between *teens* and *non-teens*. These lists of terms are built for each demographic segment.

In Step 1–2 and Step 1–4, feature vectors are extracted from user tweets. A feature vector of a user for a demographic segment is constructed based on which terms in the term list of the demographic segment appear in the tweet history of the user. Table 4 shows an example of a feature vector of users A, B, C,..., Z for a demographic segment. Feature M_{A1} is 1 in case term T_1 appears in the tweet history of user A, otherwise M_{A1} is 0. In Step 1–2 of the training phase, the field "label" is used to represent whether each user belongs to the specified demographic segment or not.

In Step 1–3, SVMs are built for each demographic segment from the feature vectors extracted in Step 1–2. SVMs are popular classifiers in the field of text categorization problems; they learn the features from training datasets. In Step 1–5, we estimate the demographic segment of unknown users. The feature vectors of a target user are extracted for each demographic segment. The SVMs output the estimated probabilities of a user for each demographic segment. The estimated probabilities are normalized so that the sum of the estimated probabilities of the demographic segments in the demographic category becomes 1. We regard the normalized estimated probability as the degree of membership in each demographic segment. The degree of membership of the text-based method Dt_i for demographic segment i is defined as shown in Eq. (5) using the estimated probability Pt_j for demographic segment j and N for the number of demographic segments in the demographic category. The target user is estimated to be in the demographic segment with the maximum degree of membership. Fig. 4 shows the detailed algorithm for demographic estimation, where the SVMs calculate estimated probabilities from the feature

Table 2
Document counts for the calculation of AIC.

	t Appears	t Does not appear	Total
D_p	$N_{11}(t)$	$N_{12}(t)$	N_p
D_n	$N_{21}(t)$	$N_{22}(t)$	N_n
Total	$N(t)$	$N(\sim t)$	N

Table 3
Examples of document counts and $E(t)$ values.

Term	$N_{11}(t)$	$N_{12}(t)$	$N_{21}(t)$	$N_{22}(t)$	$E(t)$
Examination	261	61	639	2640	2642
Beer	70	819	830	1882	−214.6
Bucket	17	51	883	2650	0

Table 4
Example of the feature vectors of a demographic segment.

	T_1	T_2	T_3	···	T_N	Label
User A	M_{A1}	M_{A2}	M_{A3}	···	M_{AN}	1
User B	M_{B1}	M_{B2}	M_{B3}	···	M_{BN}	0
···	···	···	···	···	···	···
User Z	M_{Z1}	M_{Z2}	M_{Z3}	···	M_{ZN}	0

vectors of teens, twenties, thirties, and over forties. The degree of membership is calculated from the estimation probabilities of each demographic segment. The degree of membership is also used for the purpose of combining the text-based method and the community-based method in Section 3.3.

$$Dt_i = \frac{Pt_i}{\sum_{j=1\cdots N} Pt_j} \tag{5}$$

3.2. Community-based method

As described in Fig. 3, the community-based method estimates demographic information from the follower/followee relations of unknown users. The community-based method consists of the following two steps. Step 2–1: Community extraction from follower/followee relations of unknown users, where follower and followee relationships can be regarded as directed links between two users. Step 2–2: Estimation of the demographics of the extracted communities. In the demographic estimation problem, a user belongs to multiple demographic categories such as age, gender, area, and so on. Therefore, we need to estimate the demographic category of each community group. Fig. 5 shows the details of the proposed community-based method. A user community is extracted from the followers and followees of the target user and clustered into groups. Then, the demographic category of each community group is estimated by analyzing the demographic distribution of the group members by the text-based method from the followers and followees with sufficient numbers of tweets.

In Step 2–1, communities are extracted from follower/followees of the target user. Community C is initialized with the target user. Neighbors of each user in community C are added until the number of users in the community becomes larger than the threshold T. Users with less than 200 tweets are not included in the community C because they reduce the estimated accuracy in the demographic estimation targeting the users in the clustered groups in Step 2–3.

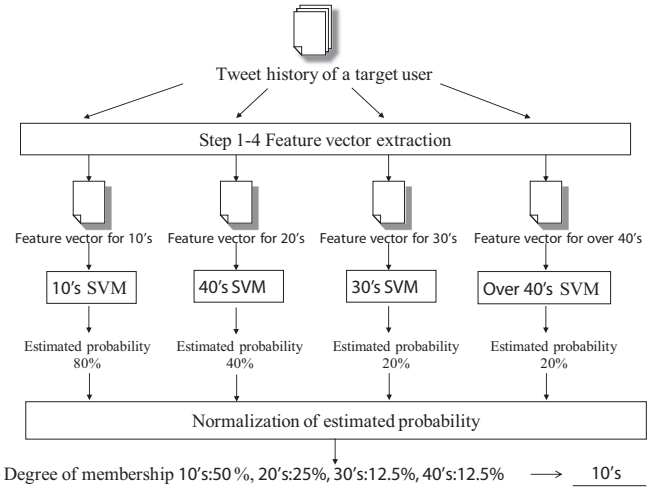

Fig. 4. Detailed algorithm of demographic estimation in Step 1–5.

In Step 2–2, the users in the extracted community C are clustered by the clustering methods for community networks. For analyzing large-scale networks, several effective methods have been proposed [23–25]. We use a tool called Fast Modularity [26], which implements the clustering algorithm of Clauset–Newman–Moore (CNM) [23]. Other methods for clustering nodes in a directed graph such as [24,25] can also be used. Fast Modularity creates several clusters with a central focus on the target user. The number of clusters, which normally lies between three and five, is automatically decided by the CNM algorithm. Fast Modularity has parameters for weighting each edge of the directed graph. In this paper, we set those parameters in a uniform value. The community extraction and clustering of Steps 2–1 and 2–2 of the proposed method is summarized in the following equation:

$$\begin{aligned}&\text{Community } C \text{ is initialized with the target user}\\&\text{While the size of } C < \text{Threshold } T\\&\qquad\text{Add 1 hop neighbors to the community}\\&\text{Clustering the community } C \text{ into multiple groups based on the CNM method}\end{aligned} \tag{6}$$

Threshold T determines the size of the community network, which is in the trade-off relationship of the accuracy and the complexity of community-based estimation. When T is too small, the extracted communities do not reflect their characteristics. On the other hand, when T is too large, many non-related users are included in the communities. In Section 4, we evaluate the accuracy of age estimation by the community-based method with the value of T being 100, 300 and 500. Follow/Follower relationships often include news groups and celebrities, which can distort the clusters due to their large-scale networks. We eliminate users who have more than 1000 followers.

In Step 2–3, the demographics of the extracted communities are estimated. We focus on the difference in the distribution of demographic segments of the clustered groups, which reflects the characteristics of members of each cluster. For example, members of a group of local friends probably have similar ages and areas. Coworkers and hobby groups have the same occupations and hobbies. We apply the text-based demographic estimation method to the users of each cluster constructed by the CNM method. For each demographic segment, the maximum ratio among each community is selected as the degree of membership. The proposed community-based method is as follows (Eq. (7)). Here, G_k is the community extracted by the CNM method. $Ratio(i, G_k)$ is the ratio of demographic segment i in community G_k. K is the number of extracted communities. Dc_i is the degree of membership of the community-based method of demographic segment i. N is the number of demographic segments in the demographic category.

In the case of age estimation, for example, a user may have a ratio of demographic segments for each community shown in Table 5. In the example, the maximum ratio is 25% of Group C for teens, 85% of Group A for twenties, 40% of Group B for thirties, and 30% of Group B for over-forties. The degree of membership is defined by normalizing the maximum ratio of each demographic segment.

$$\begin{aligned}&Max\_ratio(i) = max_{k=1\cdots K} Ratio(i, G_k)\\&Dc_i = \frac{Max\_ratio(i)}{\sum_{j=1\cdots N} Max\_ratio(j)}\end{aligned} \tag{7}$$

3.3. Hybrid of text-based and community-based method

In this section, we describe the hybrid method combining the text-based method with the community-based method (Step 3–1 of Fig. 3). Since the community-based method focuses on followers/followees independently from the target user's tweets, the hybrid of the text-based method and community-based method is

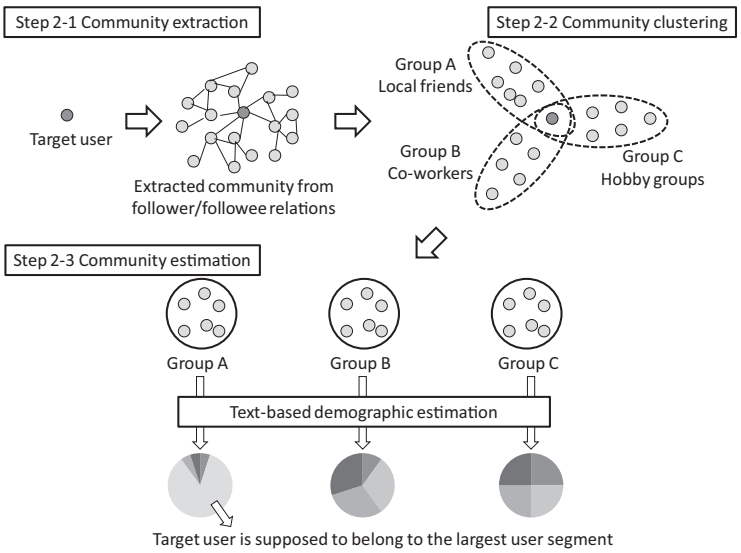

Fig. 5. Overview of the community-based demographic estimation method.

expected to be effective even for users with few tweets. We found out that the degree of membership calculated by the text-based method is high for users with plentiful text features and low for users with few text features. The community-based method can improve the estimated accuracy for users with low degrees of membership. The hybrid method that incorporates the advantages of both methods is expected to work effectively.

In the community-based method, demographic segments of a user are estimated based on the characteristic biases in the distributions of demographic segments of the follower/followees in the clustered groups. However, the community-based method cannot always estimate all demographic categories of a user. Some users only have follower/followee relations in a limited number of demographic categories. For example, a user may create follower/followee relations with users who share the same hobby as well as with co-workers, but not with local friends. In that case, the estimated accuracy of the community-based method for the user is high in the demographic categories of hobby and occupation because characteristic biases are observed in the distributions of demographic segments of the follower/followees in the clustered groups of hobby and co-workers. As for area, however, characteristic biases are not observed in the distributions of demographic segments of the follower/followees in any clustered groups of the users. For that reason, the proposed hybrid method introduces a threshold R for estimating the degree of membership calculated by the community-based method. If a user has a larger degree of membership than the threshold in a demographic segment, the user is considered to belong to the demographic segment. Then, the degree of membership of the hybrid method Dh is calculated based on the degree of membership in the text-based method Dt and community-based method Dc as shown in Eqs. 5, 7, 8. When R is too low, the estimated accuracy decreases because the estimation uses little characteristic biases. When R is too high, on the other hand, the estimation for only a limited number of users improves. We evaluate the estimated accuracy in the case of age estimation by the hybrid method in the several thresholds in Section 4.

$$Dh = \begin{cases} \frac{Dt+Dc}{2}, & \text{if } Dc > \textit{Threshold } R \\ Dt, & \text{if } Dc \leqslant \textit{Threshold } R \end{cases} \quad (8)$$

We have randomly collected 1000 Twitter users by using a streaming API. Fig. 6 shows the potential benefit of the hybrid of the text-based method and the community-based method. In case

Table 5
Example of the degree of membership of the community-based method for age.

	10s	20s	30s	Over 40s	Sum
Distribution of Group A (%)	5	85	5	5	100
Distribution of Group B (%)	5	35	40	30	100
Distribution of Group C (%)	25	25	25	25	100
Maximum ratio (%)	25	85	40	30	–
Estimated degree of membership	0.14	0.47	0.22	0.17	1

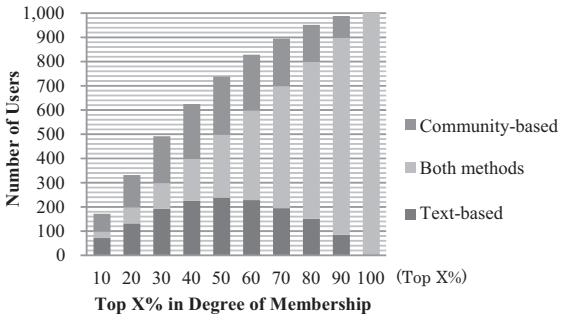

Fig. 6. Differences in the degree of membership of the text-based method and the community-based method.

there is significant crossover between users who are estimated with a high degree of membership by the text-based method and the community-based method, the hybrid of the two methods does not work effectively. In Fig. 6, the number of users in the following three sets is shown; (1) high degree of membership by the text-based method, (2) high degree of membership by the community-based method, and (3) high degree of membership by both methods. For example, when the top 10% of users (100 users of the highest degree of membership) are compared, the number of users with a high degree of membership by the both methods is 26, when either the text-based or community-based method alone yielded 74. Fig. 6 shows that the users who are given a high degree of membership differ between the text-based method and the community-based method, which means that combining these methods improves estimated accuracy.

4. Performance evaluation of demographic estimation

We evaluated the performance of the proposed methods using the datasets, evaluation metrics, and experimental environments described in the following subsections.

4.1. Collection of demographic information

On Twitter, a small number of users state their age, gender and so on in the biography field and their location information in the location field of their user profiles. We automatically collect these users with known demographics and manually label them. Although the algorithms of the proposed text-based and community-based methods are applicable to any language, we need to decide a language for the performance evaluation because the implementation includes language processing. In this paper, we evaluated the performance of the proposed method with Japanese tweets. The scalability and effects of the cultural background are discussed in Section 5.

First, we collect tweets at random using a streaming API of Twitter, but keep only the Japanese tweets by eliminating all tweets that contain no Japanese characters. There are certain writing styles relevant as demographic information. As for age, some directly state their age, such as "37-year-old," some state their year of birth, such as "born in 1974," and others simply state their age segment, such as "thirties." As for gender, "man," "gentleman," "boy," and so on, are keywords for male users, while "woman," "lady," "girl," etc., are keywords for female users. We detect this kind of information using dictionary-based keyword matching in consideration of spelling inconsistencies.

Automatically collected demographic information is not always correct in cases such as "mom of a 15-year-old boy." Therefore, the assignment of the demographic category is verified manually by two judges. In order to avoid bias in the collected demographic information, we do not use keywords that would produce indirect demographic information such as "high school" for teens or "husband/wife" for males/females.

We collect known users who state either of their age, gender, residence area, occupation, hobby or marital status from 100,000 user profiles. Table 6 shows the number of users we collect for each age segment. For example, the total number of teens is 1109 of which 326 are 18-year-old, which accounts for 29.4% of all teens. Although the numbers of younger people (less than 15 years old) and elderly people (122 users in their fifties or older are included in the age segment of over 40s) are small, the rest are collected uniformly. Six hundred users for each demographic segment are selected in the same distribution as the collected known users. Table 7 shows a summary of the number of user profiles that we use for the experiment, where genders and areas are collected in the same manner as age. The known users in each demographic segment are separated into two halves. One half is used as training dataset and the other is used as an estimation target in each demographic segment. We execute 5 repetitions of 2-fold cross-validation. Namely, we randomly separate known users into halves in each demographic segment. One of them is used for training and the other is used for test. The average performance is derived from 5 trials.

Table 6
Age distribution in collected known users.

Lower digit	Number of users (ratio of users %)			
	10s	20s	30s	Over 40s
0	4 (0.4)	306 (8.4)	160 (10.2)	80 (7.8)
1	4 (0.4)	228 (6.3)	131 (8.3)	85 (8.3)
2	4 (0.4)	354 (9.7)	118 (7.5)	60 (5.9)
3	18 (1.6)	320 (8.8)	114 (7.2)	62 (6.1)
4	18 (1.6)	330 (9.1)	88 (5.6)	27 (2.6)
5	57 (5.1)	332 (9.1)	104 (6.6)	49 (4.8)
6	123 (11.1)	306 (8.4)	77 (4.9)	31 (3.0)
7	167 (15.1)	337 (9.3)	67 (4.3)	45 (4.4)
8	326 (29.4)	209 (5.7)	54 (3.4)	24 (2.4)
9	300 (27.1)	187 (5.1)	46 (2.9)	27 (2.6)
Age segment	88 (7.9)	726 (20.0)	615 (39.1)	530 (52.0)
Total	1109	3635	1574	1020

4.2. Experimental environment and evaluation metrics

4.2.1. Experimental environment

We conducted the experiment on a terminal with a single-core 3.2-GHz processor, 8 GB of RAM and CentOS. We used LibSVM [27] as a classifier and MeCab [28] as a morphological analyzer to parse terms from Japanese tweets. The implementation was done using the programming language C.

4.2.2. Evaluation metric

We evaluated the text-based method, community-based method, and hybrid method using the following two evaluation metrics. (1) Recall, Precision and F-measure are common metrics for classification methods, which are defined as shown in Eq. (9). Recall and precision are in a trade-off relationship with each other, where the degree of membership described in Sections 3.1 and 3.2 is the parameter for tuning the trade-off. High-precision estimation obtained by tuning the parameter is effective for targeted advertising, such as sending specific information to a specific user segment. (2) The accuracy of the distribution ratio in each segment is a metric to clarify that the estimation errors are not biased towards a particular demographic segments. For marketing use, the distribution ratio of the user segments is important to perform, e.g. a follower analysis of a company. The estimated distribution ratio of the users may not be correct despite the high recall and precision when the methods have a tendency to place users into a specific demographic segment. For example, let us assume a gender estimation targeting 50 males and 50 females. In case the method estimates 50 users as males with 40 users correctly estimated and estimates 50 users as females with 40 users correctly estimated, then the

Table 7
Number of known users used in the experiment.

Demographic category	Demographic segments	# Of users in each segment	Total # of users
Gender	Male or Female	600	1200
Age group	10s, 20s, 30s, or over 40s	600	2400
Area	Hokkaido/Tohoku, Kanto, Hokushinetsu, Tokai, Kinki, Chugoku/Shikoku, or Kyushu/Okinawa	600	4200
Occupation	Employee, Part-time, Self Employed, Civil Servant, Homemaker, Student, or Without occupation	600	4200
Hobby	Reading, Gourmet, Vehicle, IT & Electronics, Games, Pets & Plants, Sports, Travel, Fashion, Music, TV & Movie, or Arts	600	7200
Marital status	Married or single	600	1200

accuracy is 80% and the distribution error is 0%. On the other hand, if the method estimates 60 users as males with 40 users correctly estimated and estimates 40 users as females with 40 users correctly estimated, then the accuracy is 80% as well, but the distribution error is significant. The error E for the distribution ratio of each demographic segment and the average error E_{avg} are defined as stated in Eq. (10). Here, T denotes the demographic, such as age, gender, or area. $T_1, T_2, \ldots, T_n$ are segments of T such as teens, twenties, or thirties, while n is the number of segments for each demographic category. $U_t(T_i)$ and $U_e(T_i)$ are the number of target users and the number of estimated users in each demographic segment T_i.

$$\text{Recall} = \frac{\text{Number of correctly estimated users}}{\text{All estimated target users}}$$

$$\text{Precision} = \frac{\text{Number of correctly estimated users}}{\text{Number of users judjed as a specific demographic segment}}$$

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{9}$$

$$E(T_i) = \frac{U_e(T_i)}{U_t(T_i)} - 1$$

$$E_{avg}(T) \frac{\sum_1^n (|U_t(T_i) - U_e(T_i)|)}{\sum_1^n U_t(T_i)} \tag{10}$$

4.3. Experimental results

Tables 8 and 9 show examples of terms extracted by the text-based method. Table 10 shows the relation between the estimated accuracy and the number of users for clustering in the community-based method. Table 11 shows the relations between estimated accuracy of the hybrid method and the value of the threshold R. Fig. 7 shows the trade-off of recall and precision. Table 12 shows the general performance of each demographic segment based on F-measure. Table 13 shows the value of E.

Table 8 shows the terms extracted in the demographic segments of age, gender, and area of residence. The extracted term lists reflect the cultural background of the users. In Japan, compulsory education, for example, is 9 years in total, 6 years of elementary school and 3 years of junior high school. After that, most people go on to high school and university. They start working in their early twenties. In Japan, from the age of 20, people have the right to vote, drink alcohol and smoke cigarettes. Because of that, terms related to school are extracted from teens' documents. Terms related to college and job-hunting are included in twenties' documents. Terms for thirties are related to work and home life. Terms for forties are related to home life, politics, and personal health. Terms for describing their partners are characteristic of both males and females. Terms related to work, politics, and home electronics are characteristic for males. Terms characteristic for females include housework and food. Terms related to the area of residence include the names of places, local transport facilities, the names of local TV stations, and terms in local dialects. Table 9 shows the terms extracted in the demographic segments of occupation, hobby, and marital status. Terms related to occupations are related to the activities of each job. As for hobbies, IT & Electronics contains the names of devices and operating systems. Fashion contains terms such as make-up and clothes. Music contains terms related to music. Terms for married users are related to family affairs. In contrast, terms for unmarried users are related to love affairs and places to go on a date.

Table 10
Relations between estimated accuracy of the community-based method and the number of users for clustering.

Category	Segment	100 users	300 users	500 users
Age	10s	50.9	64.7	63.7
	20s	35.9	48.2	41.3
	30s	43.6	41.2	48.6
	40s over	40.3	49.5	43.1
	Average	42.7	50.9	49.2

Table 11
Relations between estimated accuracy of the hybrid method and the value of the threshold R.

Category	Segment	Text-based (without hybrid)	Top 30%	Top 50%	Top 70%
Age	10s	72.6	75.1	73.2	70.4
	20s	53.5	57.5	54.9	51.2
	30s	53.8	55.8	56.0	50.3
	40s over	62.2	63.0	62.4	60.9
	Average	60.5	62.9	61.6	58.2

Table 8
Examples of extracted terms (age, gender and area).

Age				Gender		Area	
10s	20s	30s	40s over	Male	Female	Kanto	Kinki
Mathematics	University	Work	Son	Government	Husband	Shinjuku	Osaka
School	Part-time	Company	Holiday	Android	Mother	Ikebukuro	Umeda
Examination	Seminar	Business	Golf	Wife	Bath	Shibuya	Kyoto
Test	Job-hunting	Boss	Diplomacy	Company	Laundry	Yamanote-line	Yakedo (dialect)
Physical Education	Lecture	Beer	Backache	Google	Lunch	Akihabara	Hanshin

Table 9
Examples of extracted terms (occupation, hobby and marital status).

Occupation			Hobby			Marital status	
Employee	Homemaker	Student	IT & Electronics	Fashion	Music	Married	Unmarried
Office	Husband	Class	IPad	Manicure	Song	Son	Boyfriend
Work	Housework	School	IPhone	Denim	Live	Husband	Lover
Attendance	Laundry	College	Mac	Fashion	Album	Daughter	Love
Boss	Mother-in-law	Examination	OS	Makeup	Band	Home	Bored
Commuting	Mom	Anime	Android	Clothes	Music	Mam	Karaoke

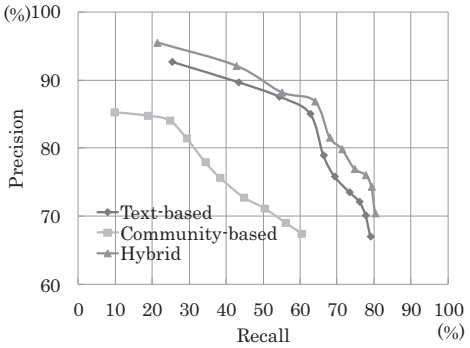

Fig. 7. Comparison of recall and precision (%) of the proposed methods for the estimation of teens.

We used the tweets of 300 known users used for term extraction. The quality of extracted terms increases with the number of known users. For the tweets of 200, 300 and 400 users respectively, we evaluated the top 2000 terms with high $E(t)$ values used for profile estimation. When increasing the number of users from 200 to 300, 719 new words appear in the top 2000. Compared to this result, only 117 new words appear in the top 2000 when increasing the number of known users from 300 to 400. From this result, we perform the evaluation with 300 known users.

As described in Section 3.2, the community-based method has a parameter, which represents the number of users for clustering T. When T is too small, the extracted communities do not reflect their characteristics. When T is too large, on the other hand, many non-related users are included in the communities. We evaluated the estimated accuracy of age estimation by the community-based method with T set to 100, 300 and 500. Table 10 shows that the estimated accuracy is lower when T is 100. On the other hand, the estimated accuracy is almost the same when T is 300 or 500. Based on this result, we set T to be 300 in the subsequent experiments.

As described in Section 3.3, the hybrid method applies a threshold R to the community-based method. When R is too small, the estimated accuracy decreases because the estimation uses little characteristic biases. When R is too large, on the other hand, the estimation is improved for only a limited number of users. We evaluated the estimated accuracy of age estimation by the hybrid method for several thresholds. R is set so that the users with the top 30%, 50% and 70% membership degrees in the community-based method are applied. Table 11 shows that the estimated accuracy is the highest when R is set to the top 30%. When R is set to the top 50%, an improvement in the estimated accuracy is observed, but it is small compared to when R is set to the top 30%. When R is set to the top 70%, the estimated accuracy of the hybrid method is lower than the text-based method only. This is probably because estimation with a low degree of membership is used. Based on these results, we set R to the top 30% in the subsequent experiments.

Fig. 7 shows the trade-off of recall and precision for teens when the users are estimated in descending order of the degree of membership such as the top 10%, 20%, 30%, ..., 100%. The precision of the text-based method and community-based method is high in the low-recall range. As described in Section 3.3, the hybrid method integrates these two methods at the threshold of degree of membership. In the case of Fig. 7, the precision of the top 30% of the text-based method is higher than for the top 10% of the community-based method. Since the community-based method works effectively in the range where the degree of membership of the text-based method is low, the threshold is set to 30% accordingly. In the same manner, the threshold is set to a point where the precision of the text-based method is higher than the precision of the top 10% of the community-based method. Looking at the column of the top 100%, the hybrid method surpasses both the text-based method and community-based method. In the hybrid method, the precision is higher than 90%, which we assume credible enough, in the range of less than 50% of the recall value. For example, people in the advertising department can send half of the teen Twitter users a commercial message about products targeting teens.

Table 12 shows the average F-measure accuracy and its standard deviation of 5 repetitions of 2-fold cross-validation for each demographic segment estimated by the text-based method, the community-based method and the hybrid method as general metrics to understand the features of each method. As for the text-based method, the estimated accuracy for teens is relatively high and the estimated accuracy for twenties and thirties is relatively low. Most teens may have similar characteristics, such as being students. On the other hand, the classification of people in their twenties and thirties is assumed to be difficult because people in their twenties and thirties may have similar lifestyles in each other. The estimated accuracy for gender is high for both males and females. The estimated accuracy for area is low in Kanto, which is the capital region of Japan. People from other regions may temporarily visit and tweet about the Kanto area. The estimated accuracy for occupation is especially high for students and homemakers, where the terms are characteristic and the variety within the communities is relatively small. The average estimated accuracy for hobbies is limited to 35.2%. Considering that the number of hobby categories is twelve, the estimation method is effective. The average estimated accuracy for material status is as high as 83.0% in the hybrid method.

As for the community-based method, although it produces F-measures lower than those from the text-based method, the precision is high in the low-recall range. The characteristic of the performance for age estimation is similar to the text-based method. As for area, its estimated accuracy for the Kanto area (Tokyo metropolitan region) is high unlike the text-based method. This is probably because the performance reduction due to temporal visitors from other areas observed in the text-based method is avoided by using community information. Application of the hybrid method increased the estimated accuracy for each demographic category. Particularly, the average estimated accuracy for occupation increased by 6.1 points in F-measure compared to the text-based method. The average estimated accuracy for age, area, and hobby also increased the F-measure by 2.4 points, 4.7 points, and 2.5 points, respectively. The hybrid method indicates better average performance compared to the text-based method in occupation, age, area and hobby with statistical significance shown in bold fonts in the table, according to a paired t-test ($p < 0.05$). Detecting communities consisting of users with the same demographic segments increases the estimated accuracy of users whose estimation is difficult only from text information. A large improvement was observed in the occupation derived from co-workers, school friends, and networks of homemakers, who are assumed to affect online communities. The accuracy for gender did not improve because the text-based method and the community-based method have similar tendency in estimation errors. For example, females work at politics or home electronics are estimated as males by the text-based method. They often have a social group which is constructed from male co-workers. Therefore, the community-based method cannot estimate them as females.

Table 13 shows the average error and its standard deviations of 5 repetitions of 2-fold cross-validation for the distribution ratio of user segments in both the text-based method and the hybrid method. The gaps between 300 target users for each demographic segment and the number of users categorized into each demo-

Table 12
Estimated accuracy of the proposed methods in F-measure (%).

Category	Segment	Text-based	Community-based	Hybrid
Age	10s	72.9 ± 0.9	63.6 ± 1.7	**75.7 ± 1.2**
	20s	54.0 ± 0.5	42.9 ± 1.1	**58.2 ± 1.0**
	30s	54.2 ± 0.7	49.4 ± 1.3	**56.2 ± 0.6**
	40s over	63.2 ± 0.7	43.5 ± 0.8	63.9 ± 0.9
	Average	61.1 ± 0.7	49.9 ± 1.2	**63.5 ± 0.9**
Gender	Male	83.4 ± 0.6	75.7 ± 1.1	84.2 ± 1.0
	Female	84.2 ± 0.8	61.8 ± 1.2	84.7 ± 0.5
	Average	83.8 ± 0.7	68.8 ± 1.1	84.5 ± 0.7
Area	Hokkaido/Tohoku	71.8 ± 0.8	51.8 ± 2.1	**78.1 ± 1.7**
	Kanto	57.9 ± 0.7	66.5 ± 1.1	**61.4 ± 0.9**
	Hokushinetsu	76.1 ± 0.9	45.2 ± 0.9	**81.0 ± 0.6**
	Tokai	71.4 ± 0.8	47.3 ± 2.5	74.6 ± 1.9
	Kinki	73.5 ± 1.5	65.1 ± 1.4	76.6 ± 0.9
	Chugoku/Shikoku	71.8 ± 0.9	54.6 ± 1.6	**79.1 ± 1.7**
	Kyushu/Okinawa	75.8 ± 1.1	60.2 ± 1.5	**80.3 ± 1.4**
	Average	71.2 ± 0.9	55.8 ± 1.6	**75.9 ± 1.3**
Occupation	Civil servant	72.3 ± 2.1	66.7 ± 1.5	74.8 ± 1.2
	Employee	64.7 ± 1.0	55.7 ± 1.9	**66.9 ± 2.0**
	Homemaker	68.2 ± 1.0	67.7 ± 1.6	**76.7 ± 1.4**
	Self employed	59.0 ± 1.8	58.2 ± 2.1	**74.4 ± 2.5**
	Part-time	64.9 ± 1.8	56.3 ± 2.3	**74.2 ± 1.0**
	Student	65.8 ± 1.3	60.3 ± 0.8	**80.4 ± 1.3**
	Without occupation	63.8 ± 1.1	61.0 ± 1.7	53.8 ± 1.2
	Average	65.5 ± 1.4	60.8 ± 1.7	**71.6 ± 1.5**
Hobby	Music	35.2 ± 0.8	34.2 ± 1.5	35.1 ± 1.0
	Reading	34.3 ± 0.9	28.6 ± 0.9	**39.7 ± 1.5**
	Games	42.7 ± 1.3	38.5 ± 1.4	39.8 ± 0.7
	TV & Movie	35.6 ± 0.8	34.8 ± 0.8	37.4 ± 1.3
	IT & Electronics	43.0 ± 0.6	41.2 ± 1.2	**47.5 ± 1.6**
	Travel	31.8 ± 0.9	32.2 ± 1.3	33.7 ± 1.4
	Gourmet	29.5 ± 0.7	32.1 ± 1.4	**34.2 ± 1.1**
	Fashion	29.5 ± 1.1	28.4 ± 1.2	30.1 ± 0.7
	Sports	26.4 ± 0.9	31.1 ± 0.8	**31.6 ± 1.1**
	Pets & Plants	33.7 ± 0.9	32.1 ± 1.6	**40.3 ± 1.8**
	Vehicle	43.5 ± 1.3	38.5 ± 1.9	44.1 ± 3.1
	Arts	36.9 ± 1.1	34.5 ± 0.7	**39.2 ± 1.0**
	Average	35.2 ± 0.9	33.9 ± 1.2	**37.7 ± 1.4**
Marital status	Married	83.3 ± 0.6	77.2 ± 1.0	83.6 ± 0.7
	Unmarried	82.7 ± 0.7	75.9 ± 1.1	83.7 ± 1.0
	Average	83.0 ± 0.6	76.6 ± 1.0	83.6 ± 0.8

graphic segment are evaluated. Considering age, for example, the number of total target users is 1200, consisting of 300 per segment. The text-based method categorized 346.6 users as teens, 263.8 users as twenties, 313.6 users as thirties, and 276.0 users as over-forties. The error E calculated by Eq. (10) in Section 4.1 is 15.5% for teens, −12.0% for twenties, 4.5% for thirties, and −8.0% for over-forties. The average error E_{avg} of the hybrid method is smaller than that of the text-based method by 2.0 points for age, 2.9 points for gender, 2.1 points for area, 10.0 points for occupation, 7.9 points for hobby, and 1.7 points for marital status. The hybrid method indicates better average performance compared to the text-based method in occupation, age, area and hobby with statistical significance shown in bold fonts in the table, according to a paired t-test ($p < 0.05$). The result shows that the hybrid method improves the estimated accuracy for the distribution of user demographics. The hybrid method is particularly effective for the categories occupation and hobby, for which the community-based method works well. With the hybrid method, the E_{avg} decreased to less than the 10% for each demographic.

Although we compare the performance of the hybrid of the text-based and the community-based methods, the hybrid of another text-based method [4–6] and the community-based method is assumed to improve performance in the same manner, because the sources used by text-based and community-based methods are independent of each other. The hybrid method required less than one second of processing time per CPU for estimating all demographics for a user. This means that only 1 month is required for estimating the demographics of all twelve million active Twitter users in Japan with a quad core computer. For real-time applications, such as analyzing opinions about on-air TV programs, it is also possible to estimate unknown users in real time.

In the performance evaluation experiments described above, we applied our methods to Twitter users who have 200 or more tweets, which is the same condition used in our previous work [19] for the purpose of the performance comparison. However, larger numbers of tweets give advantages to the text-based method. As shown in Fig. 1 of Section 2, less than half of the Twitter users have 200 or more tweets. Therefore, in order to make the contribution of the proposed hybrid method clear, we execute a performance evaluation of the estimation targeting users with the same distribution as shown in Fig. 1. We use the same model of SVM, experimental environment and evaluation metric as the experiment above. The number of target unknown users was 300 for each age segment.

Table 14 shows that the estimated accuracy of the proposed hybrid method is increased by 7.4 point in the average F-measure

Table 13
Error for distribution ratio in each demographic.

Demographic		# Of target users	Text-based		Hybrid	
Category	Segment		# Of estimated	Error (%)	# Of estimated	Error (%)
Age	10s	300	346.6 ± 3.3	15.5 ± 1.1	338.0 ± 2.8	**12.7 ± 0.9**
	20s	300	263.8 ± 2.4	−12.0 ± 0.8	276.6 ± 3.9	**−7.8 ± 1.3**
	30s	300	313.6 ± 5.3	4.5 ± 1.8	310.2 ± 5.8	3.4 ± 1.9
	40s and over	300	276.0 ± 4.3	−8.0 ± 1.4	275.2 ± 4.0	−8.3 ± 1.3
	Average			10.0 ± 1.3		**8.0 ± 1.3**
Gender	Male	300	286.6 ± 5.7	−4.5 ± 1.9	295.2 ± 3.7	−1.6 ± 1.2
	Female	300	313.4 ± 5.7	4.5 ± 1.9	304.8 ± 3.7	1.6 ± 1.2
	Average			4.5 ± 1.9		1.6 ± 1.2
Area	Hokkaido/Tohoku	300	286.2 ± 3.0	−4.6 ± 1.0	299 ± 7.6	**−0.3 ± 2.5**
	Kanto	300	352 ± 5.0	17.3 ± 1.7	331 ± 7.6	**10.3 ± 2.5**
	Hokushinetsu	300	282.4 ± 5.2	−5.9 ± 1.7	298 ± 3.6	**−0.7 ± 1.2**
	Tokai	300	286.6 ± 4.1	−4.5 ± 1.3	286 ± 3.4	−4.7 ± 1.1
	Kinki	300	301.2 ± 3.5	0.4 ± 1.2	285 ± 3.4	−5.0 ± 1.1
	Chugoku/Shikoku	300	306.8 ± 3.1	2.3 ± 1.0	309 ± 3.4	3.0 ± 1.1
	Kyushu/Okinawa	300	284.8 ± 3.6	−5.1 ± 1.2	292 ± 2.3	**−2.7 ± 0.8**
	Average			5.7 ± 1.3		**3.8 ± 1.5**
Occupation	Civil Servant	300	244.4 ± 3.8	−18.5 ± 1.3	278.8 ± 4.1	**−7.1 ± 1.4**
	Employee	300	393.8 ± 4.4	31.3 ± 1.5	298.2 ± 2.5	**−0.6 ± 0.8**
	Homemaker	300	343.8 ± 4.0	14.6 ± 1.3	338.8 ± 6.1	12.9 ± 2.0
	Self Employed	300	331 ± 6.8	10.3 ± 2.3	325.4 ± 5.3	**8.5 ± 1.8**
	Part-time	300	259.4 ± 4.3	−13.5 ± 1.4	282.6 ± 6.1	**−5.8 ± 2.0**
	Student	300	277.4 ± 3.2	−7.5 ± 1.0	289.4 ± 3.6	**−3.5 ± 1.2**
	Without occupation	300	250.2 ± 3.3	−16.6 ± 1.1	286.8 ± 4.2	**−4.4 ± 1.4**
	Average			16.1 ± 1.4		**6.1 ± 1.5**
Hobby	Music	300	267.4 ± 2.7	−10.9 ± 0.9	305.4 ± 4.3	**1.8 ± 1.4**
	Reading	300	306.2 ± 5.8	2.1 ± 1.9	259.2 ± 2.8	−13.6 ± 2.8
	Games	300	327.6 ± 3.2	9.2 ± 1.1	342.2 ± 5.6	14.1 ± 1.9
	TV & Movie	300	228.4 ± 4.1	−23.9 ± 1.4	262.6 ± 4.6	**−12.5 ± 1.5**
	IT & Electronics	300	256.8 ± 4.1	−14.4 ± 1.4	266.2 ± 3.4	**−11.3 ± 1.1**
	Travel	300	311 ± 4.7	3.7 ± 1.6	305.4 ± 5.7	**1.8 ± 1.9**
	Gourmet	300	260.4 ± 5.1	−13.2 ± 1.7	278.6 ± 3.9	**−7.1 ± 1.3**
	Fashion	300	423.2 ± 4.9	41.1 ± 1.6	348.6 ± 5.6	**16.2 ± 1.9**
	Sports	300	245.8 ± 2.0	−18.1 ± 0.7	264.2 ± 4.6	**−11.9 ± 1.5**
	Pets & Plants	300	383.2 ± 2.5	27.7 ± 0.8	340.6 ± 5.6	**13.5 ± 1.9**
	Vehicle	300	261.4 ± 2.4	−12.9 ± 0.8	298.4 ± 5.6	**−0.5 ± 1.9**
	Arts	300	328.6 ± 1.9	9.5 ± 1.9	328.6 ± 4.4	9.5 ± 1.5
	Average			16.6 ± 1.3		**8.7 ± 1.7**
Marital status	Married	300	306.8 ± 4.5	2.3 ± 1.5	301.8 ± 4.5	0.6 ± 1.5
	Unmarried	300	293.2 ± 4.5	2.3 ± 1.5	298.2 ± 4.5	−0.6 ± 1.5
	Average			2.3 ± 1.5		0.6 ± 1.5

Table 14
Estimated accuracy of the proposed methods including users with small amount of tweets.

Category	Segment	Text-based	Community-based	Hybrid
Age	10s	69.5	64.2	**74.5**
	20s	58.4	49.3	**63.6**
	30s	45.3	42.4	**52.0**
	40s over	56.5	49.0	**69.1**
	Average	57.4	51.2	**64.8**

compared to the text-based method. This result indicates that the proposed hybrid method works more effective under the practical condition that some Twitter users have small numbers of tweets.

5. Conclusions

In this paper, we proposed a hybrid demographic estimation method for Twitter users based on their tweet history and communities constructed from follower/followee relationships. There have been no previous proposals for large-scale and practical marketing analysis methods of such demographic estimation due to the difficulty of producing a method with sufficient effectiveness and accuracy for practical use.

The proposed hybrid method is applicable to multiple user demographics and to users who post few tweets by making the best use of the Twitter platform, which includes tweets as text information and followers/followees as community information. Our experimental results show that the estimated accuracy (in F-measure) of the proposed hybrid method is 84.5% for gender, 63.5% for age, and 75.9% for area. The hybrid method works effectively for the estimation of occupation, age, area and hobby. On the other hand, the hybrid method is invalid for gender. Online communities are constructed among co-workers, local friends, those who share the same hobby, etc. Those communities can easily be estimated because most of the community members have a common demographic. However, gender is mixed in most communities. Therefore, the estimated accuracy of gender does not increase with the proposed hybrid method. The performance comparison experiment of the estimation targeting users with small numbers of tweets also shows that the proposed hybrid method increases the estimated accuracy 7.4 point in average F-measure compared to the text-based method. This result indicates that

the proposed hybrid method is effective particularly in environments where some users have limited numbers of tweets. The processing time of the method is low enough to analyze all Japanese Twitter users in a month.

Demographic information attracts businesses, such as product planners, advertisement agencies, and customer support services, all of which are interested in the demographic distribution of users and their opinions. In order to deploy the technology widely, we have already implemented some novel applications for browsing online opinions in real-time, categorized by the estimated demographics and the comments' positive/negative character. We have already deployed these applications in several departments and companies. In the product planning department, for drawing up a new product proposal, our application is used to include customer opinions about the previous model as feedback for the proposal. In the customer service management department, the application is used to monitor customer opinions about released products and services. We are also working with TV stations, and the proposed technology has been used on a live debate TV program broadcast in Japan, where the question "What shrank Japan?" was announced in the official Twitter account of the program in advance and the audiences was encouraged to reply to the question. The estimated distribution of age and gender of Twitter users who answered the question was presented on the program. Remarkably, "the limitation of freedom of expression shrank Japan" was a characteristic answer among young people, and this was discussed as the topic of the program. We are also providing a demographic estimation API to a company that has developed an online opinion analysis service.

The study may have some limitations, as follows. First, as a limitation related to target languages, the proposed method is currently targeting only for Japanese tweets, which may give advantages to the proposed method compared to targeting other languages. In Japan, compulsory education, for example, is 9 years in total, 6 years of elementary school and 3 years of junior high school. After that, most people go on to high school and university. They start working in their early twenties. From the age of 20, people have the right to vote, drink alcohol and smoke cigarettes. Because of that, terms related to school are extracted from teens' documents. This cultural background affects on the terms used for demographic estimations and may differ among countries and cultures. When the proposed method targets other languages, a simple translation of the term lists is not effective. Simple approach is manually collecting known users of the target language. Then, the proposed text-based method automatically constructs term lists and estimation models. However, constructing a common term lists may be difficult or reduce the performance when the target language is used in several countries or states, where people have different cultures, lifestyles and laws. To solve these problems, by estimating the residence area of the target user, we can select appropriate term lists for the other demographics. However, the accuracy may be lower due to combining the estimations of area and other demographics. Second, as a limitation related to target users, the proposed method collects known users who clarify their demographic information such as age, gender and area on the profile field. However, those who describe their demographic information on the profile field may be relatively open to their personal information. It cannot be denied that they actively post tweets including their individual information. Questionnaire surveys for Twitter users would be necessary in order to confirm that the accuracy shown in this paper is the same for all Twitter users including unknown users. Third, as a limitation related to the performance, it is hard to improve the estimated accuracy of gender and marital status. In the text-based method, terms related to work, politics and home electronics are characteristic for males while terms related to housework and foods are characteristic for females. Since this statistical approach is applicable for most users, the accuracy of the text-based method for gender and marital status is higher than 80% by F-measure. However, it may not be major but some females are interested in work, politics and home electronics while some males are interested in housework and foods. Statistical approach does not cover these minority users. Community-based method cannot completely solve this problem, for example, females who work at politics and home electronics often have a social group which is constructed from male co-workers. Accurate estimation for these statistically minority users is our future work.

References

[1] Twitter. <http://www.twitter.com/>.
[2] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in: Proceedings of the 12th International Conference on World Wide Web (WWW 2003), 2003, pp. 519–528.
[3] J. Wiebe, E. Riloff, Finding mutual benefit between subjectivity analysis and information extraction, IEEE Transactions on Affective Computing (2011).
[4] S. Argamon, M. Koppel, J. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2) (2009) 119–123.
[5] D. Estival, T. Gaustad, S.B. Pham, W. Radford, B. Hutchinson, Author profiling for English emails, in: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING), 2007, pp. 262–272.
[6] D.D. Pham, G.B. Tran, S.B. Pham, Author profiling for Vietnamese blogs, in: Proceedings of the International Conference on Asian Language Processing (IALP), 2009, pp. 190–194.
[7] Y. Kato, D. Kawahara, K. Inui, S. Kurohashi, S. Shibata, Identifying the information sender configuration of web pages, in: Proceedings of the 2009 IEEE/ACM/WIC International Conference on Web Intelligence (WI 2009), 2009, pp. 335–340.
[8] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, Journal of the American Society for Information Science and Technology 1 (60) (2009) 9–26.
[9] A. Abbasi, H. Chen, Writeprints: a stylometric approach to identity-level identification and similarity detection, ACM Transactions on Information Systems 26 (2) (2008). No. 7.
[10] M. Koppel, J. Schler, E. Bonchek-Dokow, Measuring differentiability: unmasking pseudonymous authors, Journal of Machine Learning Research 8 (2007) 1261–1276.
[11] R. Layton, P. Watters, R. Dazeley, Authorship attribution for Twitter in 140 characters or less, in: Proceedings of the 2nd International Workshop on Cybercrime and Trustworthy Computing (CTC 2010), 2010, pp. 1–8.
[12] R.S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, B. Maia, 'twazn me!!!; ('automatic authorship analysis of micro-blogging messages, in: Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB 2011), 2011, pp. 161–168.
[13] H.G. Small, Co-citation in the scientific literature: a new measure of relationship between two documents, Journal of the American Society for Information Science 24 (4) (1973) 265–269.
[14] N.I. On, T. Boongeon, S. Garrett, C. Price, A link-based cluster ensemble approach for categorical data clustering, IEEE Transactions on Knowledge and Data Engineering (2010).
[15] P. Calado, Marco Cristo, Marcos link-based similarity measures for the classification of web documents, Journal of the American Society for Information Science and Technology 57 (2) (2006) 208–221.
[16] X. Qi, B.D. Davison, Knowing a web page by the company it keeps, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM 2006), 2006, pp. 228–237.
[17] T. Zhang, A. Popescul, B. Dom, Linear prediction models with graph regularization for web-page categorization, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 2006, pp. 821–286.
[18] X. Qi, B.D. Davison, Web page classification: features and algorithms, Journal of the ACM Computing Surveys (CSUR) 41 (2) (2009). Article 12.
[19] K. Ikeda, G. Hattori, K. Matsumoto, C. Ono, Y. Takishima, Social media visualization for TV, in: Proceedings of the International Broadcasting Convention (IBC 2011) Conference, 2011, p. D-243.
[20] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6) (1974) 716–723.
[21] Y. Sakamoto, H. Akaike, Analysis of cross-classified data by AIC, Annals of the Institute of Statistical Mathematics 30B (1) (1978) 185–197.
[22] K. Matsumoto, K. Hashimoto, Schema design for causal law mining from incomplete database, in: Proceedings of the Second International Conference on Discovery Science (DS'99), 1999, pp. 92–102.
[23] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Journal of the Physical Review E 70 (6) (2004) 066111.
[24] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Statistical properties of community structure in large social and information networks, in:

Proceedings of the 17th International Conference on World Wide Web (WWW 2008), 2008, pp. 695–704.
[25] W. Dong, M. Charikar, K. Li, Efficient K-nearest neighbor graph construction for generic similarity measures, in: Proceedings of the 20th International Conference on World Wide Web (WWW 2010), 2010, pp. 577–586.
[26] A. Clauset, Fast Modularity: Community Structure Inference Algorithm. <http://www.cs.unm.edu/aaron/research/fastmodularity.htm>.
[27] R. Fan, P. Chen, C. Lin, Working set selection using second order information for training SVM, Journal of Machine Learning Research 6 (2005) 1889–1918. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
[28] T. Kudo, K. Yamamoto, Y. Matsumoto, Applying conditional random fields to Japanese morphological analysis, in: Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), 2004, pp. 230–237.