PRISM: Profession Identification in Social Media

CUNCHAO TU, ZHIYUAN LIU, HUANBO LUAN, and MAOSONG SUN, Tsinghua University

Profession is an important social attribute of people. It plays a crucial role in commercial services such as personalized recommendation and targeted advertising. In practice, profession information is usually unavailable due to privacy and other reasons. In this article, we explore the task of identifying user professions according to their behaviors in social media. The task confronts the following challenges that make it non-trivial: how to incorporate heterogeneous information of user behaviors, how to effectively utilize both labeled and unlabeled data, and how to exploit community structure. To address these challenges, we present a framework called Profession Identification in Social Media. It takes advantage of both personal information and community structure of users in the following aspects: (1) We present a cascaded two-level classifier with heterogeneous personal features to measure the confidence of users belonging to different professions. (2) We present a multi-training process to take advantages of both labeled and unlabeled data to enhance classification performance. (3) We design a profession identification method synthetically considering the confidences from personal features and community structure. We collect a real-world dataset to conduct experiments, and experimental results demonstrate the significant effectiveness of our method compared with other baseline methods. By applying prediction on large-scale users, we also analyze characteristics of microblog users, finding that there are significant diversities among users of different professions in demographics, social network structures, and linguistic styles.

CCS Concepts: • **Social and professional topics → User characteristics;**

Additional Key Words and Phrases: Profession identification, social media, heterogeneous information, community detection

ACM Reference Format:
Cunchao Tu, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. PRISM: Profession identification in social media. ACM Trans. Intell. Syst. Technol. 8, 6, Article 81 (August 2017), 16 pages.
DOI: http://dx.doi.org/10.1145/3070665

1. INTRODUCTION

As an emerging application in social media, microblog service enables users to post messages to communicate with each other. Meanwhile, microblog users can follow each other and form social networks. Besides posting short messages and following each other, users may also contribute tags and short notes to describe themselves. The user-generated content (UGC) reserves rich facts about users, including their

This work is supported by the 973 Program (No. 2014CB340501), the National Natural Science Foundation of China (NSFC No. 61572273), and the Key Technologies Research and Development Program of China (No. 2014BAK04B03). We also thank Jiangsu Collaborative Innovation Center for Language Competence for the support to this work.
Authors' addresses: C. Tu, Z. Liu (corresponding author), H. Luan, and M. Sun, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China; emails: tucunchao@gmail.com, liuzy@tsinghua.edu.cn, luanhuanbo@gmail.com, sms@mail.tsinghua.edu.cn.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2017 ACM 2157-6904/2017/08-ART81 $15.00
DOI: http://dx.doi.org/10.1145/3070665

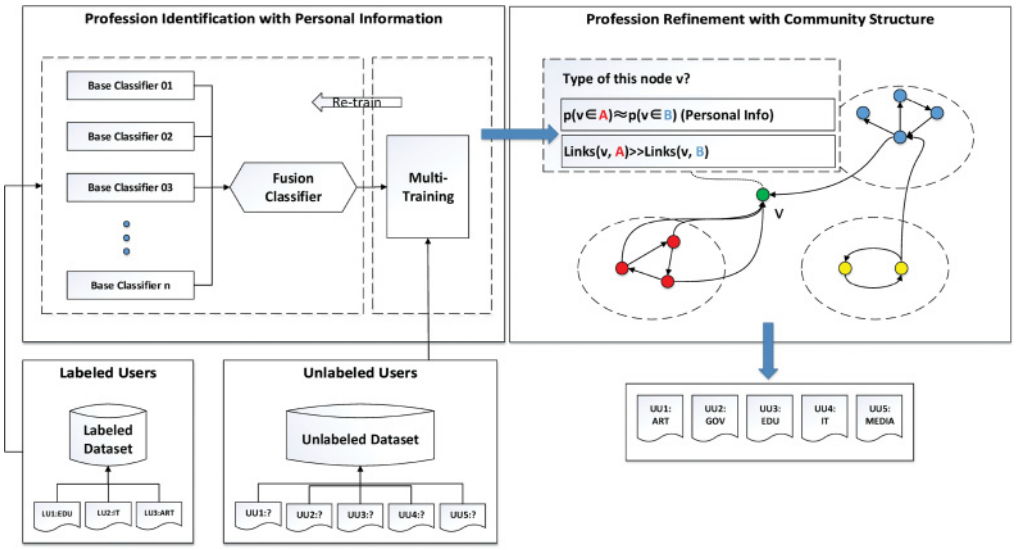

Fig. 1. The framework of PRISM.

personality traits and social attributes. Many aspects and attributes of users have been investigated based on social media data, from simple attributes such as gender and age [Burger et al. 2011] to more complicated ones such as personality [Schwartz et al. 2013], happiness [Dodds et al. 2011], and political polarity [Rao et al. 2010]. To the best of our knowledge, professions have been less investigated as a subject of prediction in social media.

According to Wikipedia,[1] profession is a vocation founded on specialized educational training and aims to supply service to others. It is also a critical social attribute of people. Sociologists have been fascinated with user professions for a long time. It is a crucial factor for many social processes and dynamics, such as social organization, social control and cohesion, differentiation and inequality, power and influence, and self- and social identity [Volti 2011]. With the development of social media, profession has become an important research subject of modern sociology. Besides benefiting research in sociology, user professions also make great contributions to commercial services such as personalized recommendations and targeted advertising. Professions of most users in social media, however, are implicit or regarded as a privacy issue. Hence, it will be beneficial for both academia and industry to effectively predict user professions based on large-scale social media data. To the best of our knowledge, user profession has been less investigated as a subject for prediction in social media. This task is the focus of this article.

The profession of a user is an essential part of human life. It may be explicitly or implicitly expressed in user-generated content in social media. Hence, user professions can be identified according to user-generated content. In this article, we take microblogs as the representative social media and explore the method of identifying user professions from microblog data.

In the context of microblog services, user professions are reflected in the following two aspects:

[1]https://en.wikipedia.org/wiki/Profession/.

(1) *Personal Information*. Microblog users provide self-descriptions and user tags and constantly post short messages. The user-generated content forms the personal information and can provide rich clues about user professions.
(2) *Network Information*. A user usually follows others to get the information he/she is interested in. The following behaviors form social networks of microblog users. In our dataset, we group users of the same profession into profession communities, which exhibit a relatively high modularity [Newman 2006] score of 0.25. This indicates strong correlations between professions and network structure and also confirms the homophily theory in sociology [McPherson et al. 2001] whereby similar users tend to form social ties.

There are several challenges making profession identification non-trivial:

(1) User-generated personal information is heterogeneous. How can we integrate this information together for identification?
(2) There are much more unlabeled users compared with those users labeled with professions. How can we effectively utilize both labeled and unlabeled data for identification?
(3) Social networks also provide strong hints for user professions. How can we take advantages of community structure and further incorporate personal information together for identification?

To address these challenges, we propose an efficient framework called **PR**ofession **I**dentification in **S**ocial **M**edia (PRISM). As shown in Figure 1, PRISM takes advantage of both personal features and community structure for user profession identification in social media.

First, for heterogeneous personal information, we present a *cascaded two-level* classifier to measure the confidence of users belonging to different professions. In the first level, we extensively extract features from different personal information sources and build separate base classifiers for each source. Afterwards, a second-level classifier integrates the classification votes and makes a final decision. Then, we further present a *multi-training* process, following the idea of co-training, to take advantage of both labeled and unlabeled users to improve classifier performance. Finally, we propose a profession identification method synthetically considering the confidence from personal features and community structure.

In the experiments, we collect more than 60,000 manually annotated microblog users from Sina Weibo (http://weibo.com), the largest microblog service in China, as our dataset. According to characteristics of microblog users, we select 14 representative professions for study, such as "art," "government," "sports," "IT," and so on. The experimental results on our dataset show that our method achieves the accuracy of 84.92%, which outperforms all other baseline methods significantly. Using this classifier, we further explore the differences on user statistics, social networks, and linguistic styles with respect to various professions.

In conclusion, our major contributions are as follows:

(1) We present three common challenges of user profiling in real-world microblog scenarios for the first time.
(2) We propose three well-focused solutions and integrate them into a unified framework to address the challenges, while most existing methods bypass them by simplifying data rather than overcoming them.
(3) We conduct profession identification experiments on a real-world dataset to verify the effectiveness of PRISM. The experimental results demonstrate that our model significantly outperforms other baseline methods. It indicates the effectiveness of PRISM on integrating heterogenous information of social media users.

(4) We investigate professional characteristics of social media users with the utilization of PRISM on a large-scale unlabeled dataset. Some interesting findings that conform to our common intuition verify the availability of our model.

2. THE FRAMEWORK

2.1. Formalizations

Let us begin with defining the problem of profession identification in social media. Suppose each user $u \in U$ in social media is represented as a bag of feature vectors $\mathcal{X}_u = \{\mathbf{x}_{u,r}\}$, where U is the user set and each $\mathbf{x}_{u,r}$ denotes a feature vector obtained from a distinct information source $r \in \{1, \ldots, R\}$. Here, R is the number of information sources. There is also a social network $G = (U, E)$, where E are connections between these users, that is, $E \subset U \times U$. Besides, we also have a set of annotated user-profession pairs $\{(\mathcal{X}_u, y_u)\}$ for training, where $y_u \in \{1, \ldots, K\}$ and K is the number of professions. Profession identification for social media users aims to classify unlabeled users into predefined professions according to their heterogenous information, including text content and network structure.

With the above definitions, we design the framework of our model as a two-step process:

(1) We represent each user as multiple feature vectors extracted from various personal information sources and build a cascaded two-level classifier to identify their professions. Furthermore, we introduce a multi-training process to improve classification performance by incorporating unlabeled data for training.
(2) We further take advantage of profession community structure to refine profession identification. We introduce the details of our method as follows.

In the following part, we introduce each step of our model in detail.

2.2. Profession Identification with Personal Information

We build a cascaded two-level classifier for profession identification.

(1) **Base Classifier Construction.** For each information source r, we build a base classifier $f_r(\cdot)$ with a set of user-profession pairs $\{(\mathbf{x}_{u,r}, y_u)\}$. With these base classifiers, for a user u and its feature vector $\mathcal{X}_u$, we can obtain an identification matrix $\mathcal{P}_u = \{p_{k,r}\}$, where $p_{k,r} = \Pr(k|\mathbf{x}_{u,r}) = f_r(\mathbf{x}_{u,r}, k)$, indicating the confidence score for categorizing user u into profession k based on information source r.
(2) **Base Classifier Fusion.** We take identified results $\mathcal{P}_u$ obtained in (1) as input features and construct a new set of user-profession pairs $\{(\mathcal{P}_u, y_u)\}$. Using these pairs, we build a fusion classifier $g(\cdot)$. The fusion classifier will assign a weight for each base classifier learned in Step 1 and fuse their identification results into the final identification scores, $\Pr(k|\mathcal{P}_u) = g(\mathcal{P}_u, k)$. We can then select the most confident label $\hat{y}_u = \text{argmax}_k \Pr(k|\mathcal{P}_u)$ as the identified profession.

2.2.1. Feature Design and Base Classifier Construction. In social media, a user generates various types of content. We take, for example, "Kai-Fu Lee," a famous Chinese IT activist. He provides a short self-description, "CEO of Innovation Works," gives some user tags such as "venture capital," "innovation works," "education," "technology," and "e-business," and also has the verification information "Chairman and CEO of Innovation works." He also has posted thousands of messages, containing rich information including words, mentioned users, URLs, entities, and hashtags. This information should be handled separately due to its distinct characteristics. In this article, we consider eight distinct sources of user-generated personal information to build features for base classifiers, which are listed in Table I.

Table I. Personal Information Sources

No.	Name	Source Description
1	DES	Self-descriptions provided by user.
2	TAG	User tags provided by user.
3	VER	Verification information for user.
4	MSG	Messages posted by user.
5	MEN	Mentioned user IDs in messages.
6	URL	URLs in messages.
7	ENT	Named entities in messages.
8	HAS	Hashtags in messages.

Among these feature sources, the features in DES, VER, and MSG are words extracted from text following the bag-of-words assumption. For TAG and HAS, we use tags as features. Besides using words in messages as features, we also extract user IDs identified by "@" in microblog messages as features of MEN, regard URLs in messages as features of URL that are usually in the form of tiny URLs [Antoniades et al. 2011], and use named entity recognition (NER) tools to extract entities from messages as features of ENT.

For each feature source, there are tens of thousands of feature candidates. We have to perform feature selection to downsize feature sets. Following the valid experience in feature selection for text classification [Yang and Pedersen 1997; Forman 2003], we use the χ^2 statistic to select representative features for each feature source. Afterwards, we build base linear classifiers for each feature source.

2.2.2. Base Classifier Fusion. The prediction result $\mathcal{P}_u$ obtained from base classifiers for user u is a matrix, which cannot be directly used as the input of fusion classifier. We concatenate the transfer matrix $\mathcal{P}_u$ into a feature and feed it into fusion classifier, that is, building a feature vector $\mathbf{z}_u$ simply by concatenating column vectors of $\mathcal{P}_u$, that is, $z_{u,k+K\times(r-1)} = p_{k,r}$. The vector size of $\mathbf{z}_u$ is $K \times R$. We can also select the maximal scores or sum up scores of each row in the prediction matrix to build a feature vector. However, in experiments, we find that the concatenation scheme significantly outperforms the other schemes (max and sum), and hence we only report the concatenation results.

We select Liblinear [Fan et al. 2008][2] to build base classifiers and fusion classifier. In this package, we select the method of L2-regularized logistic regression (LR), which is also the default setting of Liblinear. We have compared LR with support vector machine (SVM),[3] and LR performs better in both effectiveness and efficiency. Hence, in the following, we only show the results obtained with LR.

2.2.3. Multi-Training with Labeled and Unlabeled Data. In the real world, there is a much larger set of unlabeled users with no profession information. Here we want to employ the idea of co-training to perform multi-training of profession classification with both labeled and unlabeled data.

The basic idea is, after building base classifiers, that we will use them to identify professions for unlabeled users. We select the users where more than *half* the base classifiers agree on their professions and put these users with corresponding identified profession labels into a training set. Then we re-train these base classifiers.

[2]In this article, we use the Java version of Liblinear, developed by Benedikt Waldvogel, which can be accessed via http://www.bwaldvogel.de/liblinear-java/.

[3]We select LibSVM [Chang and Lin 2011] as the implementation of SVM, which can be accessed via http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

We can conduct the procedure iteratively until convergence. Multi-training is expected to enrich training data and improve classification performance with respect to both accuracy and generalization.

2.3. Profession Refinement with Community Structure

We observe from our dataset that users of the same professions tend to be friends and form communities in social networks, which is consistent with our intuition and sociology theory [McPherson et al. 2001]. Following Mislove et al. [2010], we assume that users with the same profession tend to form a profession-specific community. A relatively high modularity [Newman 2006] score of 0.25 obtained on our dataset confirms the assumption. Based on the observation, we take community structure into consideration to refine the identification results based on personal information.

Community-based profession refinement is formalized as follows. Suppose we have a social network $G = (U, E)$ and a subset of users who have profession labels and form communities for each profession, denoted as $G_k = (U_k, E_k)$ for profession k, where U_k is the set of all users of profession k and E_k is the set of edges between users in U_k. Afterwards, given a subset of users V with no profession labels, the task aims to extend existing communities by putting users from V into correct communities, that is, assigning correct profession labels, according to the effect on community quality if users are involved in.

For community-based profession refinement, it is important to define an appropriate measure of community quality for each profession-based community G_k. The community quality can be verified from two aspects, including network structure and content information, that is, *structure quality* and *content quality*.

2.3.1. Structure Quality. Structure quality measures the significance of a community from the perspective of network structure. It is intuitive that the users with the same profession will form a dense and compact profession-based community.

To formally define structure quality, we give some definitions as follows. Take G_k, the community of profession k, for example; we define $U_{\neg k} = U \backslash U_k$. We also define $E_{k,\neg k}$ as the number of links between U_k and $U_{\neg k}$ and so do $E_{k,k}$ and $E_{\neg k,\neg k}$. We also have $E_k = E_{k,k} + E_{k,\neg k}$ and $E_{\neg k} = E_{\neg k,\neg k} + E_{\neg k,k}$.

Based on the above definitions, the structure quality of G_k is formalized as

$$Q_{structure}(G_k) = \frac{E_{k,k}}{E_{k,k} + E_{k,\neg k}} - \frac{E_k E_k}{E_k E_k + E_k E_{\neg k}}, \tag{1}$$

where the first entry indicates the proportion of how many links starting from U_k are connected within the community, and the second entry is that of the corresponding random graph. $Q_{structure}$ ranges of $[-1, +1]$, and a strongly positive score indicates there is significant community structure in G_k.

This measure is originally proposed by Mislove et al. [2010] to compute the quality of a community, called *normalized conductance*. In this article, we integrate it with content quality together for profession refinement.

2.3.2. Content Quality. Content quality measures the significance of a community based on personal confidences of all users assigned in this community. In this article, we employ identification confidences from our cascaded two-level classifier to measure content quality.

We define the content quality of a community as the average confidence scores over all users in this community, denoted as $Q_{content}(G_k)$. With content quality, the algorithm can take identification results based on personal information as input for refinement.

Table II. Ratios of Professions in the Annotated Dataset (%)

No.	Category	Ratio	No.	Category	Ratio
1	media	25.6	8	education	4.0
2	government	15.1	9	fashion	3.9
3	entertainment	8.8	10	games	3.8
4	estate	8.2	11	literature	3.4
5	finance	7.0	12	services	3.4
6	IT	6.4	13	art	3.1
7	sports	5.6	14	healthcare	1.7

2.3.3. Profession Refinement. Afterwards, the overall community quality of G_k is defined as a combination of $Q_{structure}$ and $Q_{content}$,

$$Q(G_k) = \lambda Q_{structure}(G_k) + (1 - \lambda)Q_{content}(G_k), \tag{2}$$

where λ is a harmonic smoothing factor. When $\lambda = 1$, the quality measure is identical to that in Mislove et al. [2010].

With the measure $Q(\cdot)$, we conduct a greedy community extension as follows. Given a profession k, for each user $u \in V$ we compute

$$\Delta Q(u) = Q(G_k + u) - Q(G_k), \tag{3}$$

We find the user $\hat{u} = \arg\max_u \Delta Q(u)$, put $\hat{u}$ in U_k, and repeat the procedure until convergence.

After the community extension process, every unlabeled user is put into a profession community, in which users have the most similar personal information and close connections. We take the types of matched communities as the final identified profession of unlabeled users.

2.4. Complexity Analysis

In PRISM, the complexity of training LR classifier for the rth feature source is $O(K|U| \cdot |\mathbf{x}_r|t_{LR})$, where t_{LR} indicates the average computing time for an individual feature value and training example. Therefore, the training complexity of the cascaded classifier is $O(K|U|t_{LR}\sum_{r=1}^{R}|\mathbf{x}_r|)$. The community refinement complexity is $O(|V|^2K \cdot \frac{|E|}{|V|}) = O(K|V||E|)$. Suppose we iterate the multi-training for m times; then the overall complexity of PRISM is $O(mK|U|t_{LR}\sum_{r=1}^{R}|\mathbf{x}_r| + K|V||E|)$. That means that the complexity of PRISM equals the summation of training mK LR classification models and a conductance-based community detection model.

3. EXPERIMENTS AND ANALYSIS

We collect 62,415 active and influential users from Sina Weibo. These users are all verified and categorized into 14 professions by officials of Sina Weibo, known as Hall of Fame in Weib.[4] In addition, we also crawl the profiles and more than 10 million posted messages of these users with APIs to build classifiers. Moreover, we collect an additional 150,000 verified users with no profession annotations for multi-training. The ratios of various professions among these users are shown in Table II.

From the profession composition of these labeled users, we find that the users in "media" and "government" are dominant. The reasons may be as follows: (1) As the largest public social media service in China, public events are heavily discussed on Sina Weibo. Therefore, many people working in newspapers, news agencies, and social media are active here. (2) The Government of China encourages their officials to go

[4]http://verified.weibo.com/.

online and contact with citizens officially. Therefore, many national and local officials have registered in Sina Weibo.

Note that each user in our labeled dataset has only one profession. Thus, we treat the profession identification task as a single-class classification problem. In real world, users may have other profession or multiple professions, we can overcome it by setting a threshold over classification weights or introducing an "other" profession when building the labeled dataset. Besides, we only collect verified users to get credible labeled data. However, verification information is one part ofeight feature sources in our framework. As with other feature sources, some verified users also have no verification information. Our framework can adapt to the situation where this is a lack of feature sources, which is general in our dataset and actual scenarios.

3.1. Experimental Results on Profession Identification

We randomly divide the 62,415 labeled users into a training set and test set, of which 4/5 is for training and 1/5 for test. For the test set, we regard the labeled profession as the gold standard.

We select accuracy, macro-averaging precision/recall/F-Measure as evaluation metrics. Suppose the number of test users is U_{test}. If we get correct identification for U_{correct} users, then the accuracy is computed as $\frac{|U_{\text{correct}}|}{|U_{\text{test}}|}$. Accuracy evaluates per-user decisions across profession classes globally and thus is micro-averaging, whereas macro-averaging first calculates precision/recall for each profession class. That is, precision of profession k is $\frac{|U_{k,\text{correct}}|}{|U_{k,\text{predict}}|}$ and recall is $\frac{|U_{k,\text{correct}}|}{|U_k|}$, where U_k is the user set of profession k, and $U_{k,\text{predict}}$ is the user set that are predicted as profession k. And then it takes the average of these scores as overall precision P and recall R and further calculates F-measure as $\frac{2PR}{P+R}$.

As mentioned, user profession has been less investigated as a subject for prediction in social media. In particular, there are few ensemble methods that can be directly applied to the user profession identification according to the heterogenous feature sources in Sina Weibo. Thus, we adopt classifiers with single-source feature vectors and a concatenated feature vector as baselines.

3.1.1. Parameter Settings. For feature selection of base classifiers, we evaluate performance with different numbers of features and select 2,300 features for DES, 3,800 features for TAG, 4,000 features for VER, 6,600 features for MSG, 3,200 features for MEN, 2,700 features for URL, 3,600 features for ENT, and 4,100 features for HAS, which achieve the best performance for each base classifier. For the harmonic smoothing factor λ in 2, we test the performance with different settings and choose $\lambda = 0.2$ when our fusion model achieve the best performance.

3.1.2. Profession Identification with Personal Information. Table III shows the evaluation results on profession identification with various features of their combinations. In this table, the line of "Single Vector" is the baseline that represents a user by concatenating all features from multiple sources into a single vector, "Fusion" indicates the method of our cascaded two-level classifier, and "Fusion + MT" indicates the results after multi-training. From Table III, we observe that:

(1) The fusion classifier performs much better than "Single Vector." This indicates that the design of cascaded two-level classifier is necessary and efficient for integrating heterogeneous feature sources.
(2) The base classifier using VER as a feature source achieves the best performance among all base classifiers. This is consistent with the fact that verification descriptions are more informational and less noisy compared to other feature sources.

Table III. Evaluation Results for Various Features and Combinations (%)

Method	Accuracy	Precision	Recall	F
DES	31.25	51.82	28.90	37.11
TAG	38.11	50.55	31.04	38.46
VER	78.63	75.73	74.89	75.31
MSG	47.47	49.58	42.79	45.93
MEN	38.22	42.85	30.59	35.70
URL	26.38	36.47	13.68	19.90
ENT	33.86	36.88	26.95	31.15
HAS	30.91	37.44	17.60	23.94
Single Vector	39.25	48.33	34.92	40.54
Fusion	81.25	79.60	76.27	77.90
Fusion+MT	**83.38**	**82.24**	**81.35**	**81.79**

Table IV. Evaluation Results of Profession Refinement with Community Structure (%)

Method	Accuracy	Precision	Recall	F
LPA	58.86	57.05	54.53	55.76
CD	64.20	65.11	60.78	62.87
PRISM				
$\lambda = 0.1$	84.17	83.15	81.62	82.37
$\lambda = 0.2$	**84.92**	**83.78**	**81.89**	**82.82**
$\lambda = 0.3$	81.12	79.10	77.42	78.25
$\lambda = 0.5$	77.56	76.53	75.08	75.79

(3) The fusion classifier achieves much better performance compared to all base classifiers. This indicates that the fusion of base classifiers can effectively incorporate heterogenous information and significantly improve identification performance. In contrast, the simple concatenation of feature vectors may bring in noisy features and result in the decrease of performance.
(4) The accuracy and macro-averaging precision/recall/F-measure of fusion classifier with multi-training process are all larger than 80%. This indicates the identification capability of our classifier is balanced among various professions.

3.1.3. Profession Refinement with Community Structure. To evaluate the performance of our profession refinement with community structure, we take two community-based methods as baselines, that is, label propagation algorithm (LPA) and community detection (CD). LPA addresses the task as graph-based semi-supervised learning [Zhu and Goldberg 2009]. The basic idea of LPA is that the labels of a user are dependent on his or her neighbors'. By propagating labels from annotated users to unannounced users through a social network, LPA can identify the profession labels of users. As previously introduced, CD is the user profiling algorithm proposed in Mislove et al. [2010]. Both methods only consider community structure to classify users.

We show the evaluation results in Table IV. From the table, we observe that:

(1) Profession refinement with community structure achieves considerable improvement as compared to the two-level classifier. This indicates that the community structure can also provide supplementary information for profession identification beyond personal information. According to our statistical result, the lack of personal information is very common (e.g., 153 users have no verification information) and will result in the poor performance of personal information-based

Table V. Evaluation Results for Each Profession (%)

No.	Profession	Precision	Recall	F
1	media	84.04	90.60	87.20
2	government	94.03	93.78	93.90
3	entertainment	84.78	82.25	83.49
4	estate	88.22	86.92	87.57
5	**finance**	68.86	73.05	70.90
6	**IT**	72.93	68.38	70.58
7	sports	94.05	92.84	93.44
8	**education**	76.88	73.80	75.31
9	fashion	84.84	78.94	81.78
10	game	85.47	84.19	84.82
11	literature	84.68	75.99	80.10
12	**service**	65.32	57.45	61.13
13	**art**	76.84	69.92	73.22
14	healthcare	87.10	87.50	87.30

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	91.05	0.66	1.88	0.81	1.05	1.43	0.51	0.30	0.24	0.30	0.39	0.95	0.42	0.03
2	2.64	93.45	0.17	0.39	0.84	0.28	0.03	0.78	0.02	0.00	0.05	0.17	0.39	0.78
3	9.72	0.47	81.34	0.19	1.43	0.66	0.28	0.19	2.09	0.57	1.15	0.38	1.53	0.00
4	5.31	0.40	0.26	82.76	5.82	1.04	0.40	0.26	0.51	0.00	0.26	2.71	0.26	0.00
5	3.95	2.22	0.29	3.56	77.37	6.55	0.29	1.92	0.77	0.29	0.29	2.03	0.39	0.09
6	7.78	0.44	0.00	2.52	9.42	72.18	0.33	0.88	0.55	4.06	0.11	1.43	0.11	0.22
7	3.45	0.36	0.36	0.36	0.79	0.00	94.05	0.24	0.00	0.12	0.00	4.00	0.24	0.00
8	3.89	3.05	0.85	0.68	4.40	1.19	0.34	77.82	0.34	0.00	0.85	3.72	1.19	1.69
9	5.93	0.18	4.14	0.54	2.88	0.90	0.36	0.72	81.31	1.26	0.18	0.90	0.72	0.00
10	3.19	0.00	1.60	0.18	0.00	5.14	0.71	0.00	1.60	86.70	0.35	0.18	0.35	0.00
11	9.33	1.87	1.65	0.00	1.45	0.00	0.00	2.28	0.20	0.83	78.65	0.00	3.32	0.42
12	13.23	1.04	1.86	4.34	6.41	4.76	0.00	3.72	1.04	0.83	0.00	62.56	0.21	0.00
13	4.71	2.35	4.71	0.00	2.35	0.00	0.79	2.95	0.79	0.79	3.33	0.20	77.05	0.00
14	2.64	0.76	0.37	0.00	2.64	0.37	0.00	1.13	0.75	0.00	1.51	0.00	0.01	89.82

Fig. 2. Distribution of identified professions for each profession.

classification. Therefore, the consideration of community structure is beneficial to overcome such situations.

(2) Profession refinement also outperforms two community-based baselines significantly. We can see that PRISM achieves the best result when $\lambda = 0.2$. This indicates the effectiveness of personal information for profession identification. Here the community structure does not play a critical role for profession refinement compared with personal information, because in Sina Weibo personal information is much richer. The effect of social networks may be emphasized in other scenarios with richer social structure information. Note that we can tune the parameter λ by cross-validation or setting a development set in practice.

3.1.4. Error Analysis. We also show the identification performance with respect to each profession in Table V, including precision, recall, and F-measure. From the table, we can find that most professions achieve good identification performance, except for several professions, including "services," "IT," "finance," "art," and "education" (marked in bold).

To investigate the reason for these identification errors, we show the distribution of identified professions for each profession in Figure 2. In this figure, we define the entry of row-i and column-j as the ratio of the users in profession i being identified as

Table VI. User Statistics in Each Profession

No.	Profession	Gender	Message	Follower
1	media	1.25	1,776	18,555
2	government	2.66	1,270	17,724
3	entertainment	1.57	1,367	59,709
4	estate	2.42	1,375	6,661
5	finance	2.60	1,457	15,990
6	IT	3.19	1,558	17,224
7	sports	2.51	1,380	36,813
8	education	1.88	1,384	11,216
9	fashion	0.77	1,392	24,867
10	games	1.75	1,358	20,354
11	literature	0.94	1,924	31,522
12	services	2.04	1,512	7,964
13	art	2.68	1,667	14,447
14	healthcare	1.38	1,386	14,726

profession j, that is,

$$e_{ij} = \frac{\sum_{u \in U_i} k_u = j}{|U_i|}, \tag{4}$$

where k_u indicates the identification result for the user u. To make the distribution comprehensive, we also illustrate the ratio in each entry using different shades of color. From this figure, we observe that:

(1) The profession "service" tends to be categorized into "media" by mistake, and the professions "art" and "education" are usually categorized into other professions incorrectly. The reason is that there are more overlaps between these related professions, which makes the boundary between professions not so clear for identification. For example, the profession "service" usually interacts with "media," because they both involve in advertising and marketing.
(2) The professions "finance" and "IT" are usually categorized into each other incorrectly. We carry out extensive case studies and find that many top executives of companies usually have experiences in both "IT" and other business fields such as "finance," which cannot be well dealt with. The truth is also reflected in their friend network. In future, we may find more insight features to address these issues.

3.2. Profession Analysis

To explore the professional characteristics of microblog users, we utilize our model to identify the profession attributes of unlabeled users, as the statistical results on a large dataset may be more credible than the limited labeled dataset. Some interesting findings that conform to our intuition verify the availability of our model in real-world datasets.

3.2.1. User Statistics. We demonstrate user statistics of various professions in Table VI. The column "Gender" indicates the gender ratio ($\frac{\#\text{male}}{\#\text{female}}$), "Message" indicates the average number of short messages per user, and "Follower" indicates the average number of following users by each user in the profession.

It is apparent to observe from Table VI that:

(1) There is significant gender bias among professions due to professional characteristics. The professions such as "IT," "art," "government," "finance," and "sports" have a ratio larger than 2.50, meanwhile only the profession "fashion" has a ratio lower

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.00	0.00	1.00	0.00	0.15	0.09	0.04	0.26	0.43	0.05	0.93	0.14	0.60	0.30
2	0.00	1.00	0.99	0.00	0.18	0.04	0.12	0.44	0.15	0.06	0.85	0.09	0.61	0.39
3	0.00	0.00	1.00	0.00	0.01	0.01	0.02	0.08	0.89	0.06	0.65	0.09	0.82	0.21
4	0.00	0.00	0.50	1.00	0.91	0.05	0.02	0.16	0.12	0.04	0.58	0.19	0.38	0.23
5	0.00	0.00	0.73	0.10	1.00	0.98	0.03	0.51	0.24	0.06	0.77	0.21	0.56	0.33
6	0.00	0.00	0.21	0.01	1.00	1.00	0.01	0.33	0.10	0.28	0.68	0.26	0.35	0.28
7	0.00	0.00	0.99	0.00	0.01	0.01	1.00	0.06	0.25	0.04	0.19	0.06	0.29	0.20
8	0.01	0.00	0.94	0.01	0.97	0.54	0.03	1.00	0.17	0.06	0.98	0.24	0.85	0.44
9	0.00	0.00	1.00	0.00	0.04	0.02	0.02	0.07	1.00	0.07	0.41	0.16	0.94	0.25
10	0.00	0.00	0.98	0.00	0.10	1.00	0.02	0.07	0.47	1.00	0.76	0.10	0.44	0.19
11	0.00	0.00	0.98	0.00	0.04	0.10	0.01	0.41	0.10	0.27	1.00	0.10	0.94	0.31
12	0.01	0.00	0.99	0.27	0.94	0.96	0.02	0.79	0.51	0.10	0.85	1.00	0.69	0.34
13	0.00	0.00	1.00	0.00	0.07	0.03	0.02	0.27	0.92	0.10	0.98	0.14	1.00	0.24
14	0.00	0.00	0.91	0.00	0.32	0.07	0.04	0.66	0.19	0.05	0.90	0.09	0.55	1.00

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.00	0.00	0.07	0.00	0.04	0.07	0.02	0.14	0.25	0.06	0.47	0.25	0.30	0.24
2	0.00	1.00	0.00	0.00	0.02	0.02	0.01	0.30	0.04	0.05	0.13	0.19	0.18	0.31
3	0.01	0.00	1.00	0.00	0.01	0.01	0.01	0.08	0.91	0.07	0.32	0.13	0.82	0.20
4	0.00	0.00	0.00	1.00	0.08	0.03	0.01	0.05	0.05	0.04	0.06	0.44	0.09	0.19
5	0.00	0.00	0.01	0.08	1.00	0.99	0.02	0.44	0.24	0.09	0.18	0.59	0.28	0.34
6	0.00	0.00	0.00	0.00	0.98	1.00	0.01	0.21	0.06	0.62	0.17	0.60	0.14	0.24
7	0.00	0.00	0.00	0.00	0.01	0.01	1.00	0.04	0.08	0.04	0.05	0.07	0.10	0.18
8	0.01	0.00	0.01	0.00	0.76	0.54	0.01	1.00	0.10	0.10	0.61	0.67	0.71	0.46
9	0.01	0.00	1.00	0.00	0.03	0.02	0.02	0.08	1.00	0.10	0.11	0.28	0.93	0.24
10	0.00	0.00	0.00	0.00	0.01	0.99	0.01	0.04	0.15	1.00	0.29	0.16	0.19	0.15
11	0.01	0.00	0.09	0.00	0.02	0.10	0.01	0.35	0.07	0.22	1.00	0.13	0.93	0.28
12	0.04	0.00	0.01	0.65	0.59	0.98	0.01	0.77	0.27	0.20	0.18	1.00	0.32	0.31
13	0.01	0.00	0.97	0.00	0.04	0.02	0.01	0.29	0.95	0.12	0.92	0.20	1.00	0.23
14	0.00	0.00	0.00	0.00	0.11	0.07	0.01	0.56	0.13	0.05	0.29	0.15	0.18	1.00

Fig. 3. Profession affinities in the following network (top) and friend network (bottom).

than 0.80. This phenomenon verifies some theories in professional segregation that females tend to work in a narrower range of professions than males [Chafetz 1988].

(2) The average numbers of short messages are similar among professions, whereas the average numbers of followers are relatively larger for some more public professions, such as "entertainment," "sports," "literature," and "fashion," because there are more celebrities in these professions.

3.2.2. Professions and Social Networks. As we previously mentioned, professions of users in social media exhibit high correlations with social network structure. Here we define *profession affinity* to quantify the tendency of users in one profession making friends with another profession, $a(i, j) = g(100 \times (\frac{N_{i,j}}{N_i} - \frac{U_j}{U}))$, where $g()$ is the sigmoid function, N_i is the average number of friends (neighbors) for each user in profession i, among these friends $N_{i,j}$ indicates the average number of them being in profession j, and $\frac{U_k}{U}$ is the global profession distribution in the dataset. A higher score of $t(i, j)$ indicates that users in i have a higher preference to make friends with users in j. In Figure 3, we show profession affinities between professions using both numbers and shades of color on the following network and the friend network, respectively.

From Figure 3 we observe that:

(1) People tend to *follow* users in more diverse professions and make more friends in the same profession.

Table VII. Ratios of Conjunctions, Interjections, and Modal Particles Used by Users of Various Professions (%)

No.	Profession	Conj.	Interj.	M.P.
1	media	1.19▽	0.22△	2.16△
2	government	1.29	0.17	1.70
3	entertainment	1.08▽	0.26△	2.38△
4	estate	1.26	0.15	1.72
5	finance	1.39△	0.15▽	1.65▽
6	IT	1.35△	0.15▽	1.66
7	sports	1.04▽	0.25△	2.60△
8	education	1.42△	0.16▽	1.55▽
9	fashion	1.25	0.22	1.95
10	games	1.34	0.16	1.26▽
11	literature	1.31	0.27△	2.25
12	services	1.29	0.18	1.94
13	art	1.11▽	0.22△	2.06△
14	healthcare	1.76△	0.11▽	1.15▽

(2) The professions such as "entertainment," "education," and "literature" tend to be followed by many other professions; while the professions such as "media," "government," and "estate" are on the contrary.
(3) There are strong correlations between "IT" and "finance"/"game"/"service," and between "entertainment" and "fashion"/"art." This reflects the natural cooperation among professions in society.
(4) The affinities between two professions are sometimes asymmetric. For example, the users in "education" are more likely to make friends with or follow those in "art" as compared to the reverse direction.

3.2.3. Professions and Linguistic Styles. Sociolinguists believe and verify that personal attributes can be reflected via linguistic styles [Niederhoffer and Pennebaker 2002; Mairesse et al. 2007]. It is straightforward that users in different professions have their own terminologies, which are usually notional words such as nouns, verbs, and adjectives.[5] Here, we focus on the other side, that is, function words, and explore the relations between linguistic styles and professions. Although the size of function words vocabulary is relatively much smaller than notional words, function words are heavily used in daily writing and conversation.

We perform part-of-speech tagging for short messages posted by users, and in Table VII we show the ratios of conjunctions, interjections, and modal particles used by users in various professions. For each column, we highlight Top-4 professions with △ and Bottom-4 professions with ▽.

We can see that the users in "healthcare," "education," "finance," and "IT" use more conjunctions but fewer interjections and modal particles. This is consistent with the fact that most of these professions require humans apt to be more rational, logical, and precise. On the contrary, the users in "sports," "entertainment," "art," and "media" use fewer conjunctions but more interjections and modal particles. The reason may be that these four professions attract and train humans to be more emotional, imaginative, and freestyle. This preliminarily but significantly confirms the theory in sociolinguistics that personality characteristics of various professions may be reflected via their

[5]For this reason, many Input Method Editors provide terminology dictionaries for users.

language usage styles. Of course, this is only an initial observation, which deserves deeper investigation in future.

4. RELATED WORK

4.1. Profession and Sociology

Profession is an important personal attribute, playing a crucial factor in many social processes. Profession attracts sociologists as a distinct type of social organizations, focusing on key topics such as the division of labor, profession communities, and social activity of work [Rothman 1987; Volti 2011]. Sociologists are also interested in relations between professions and other attributes such as personality [Holland 1997].

Similarly to other fields in sociology, the data for traditional profession research are primarily collected via inquiry or survey, and thus data scalability is severely limited. As the development of social media, it provides rich information of user behaviors for sociology research. Hence, data-driven computational social science emerges [Lazer et al. 2009] and has achievements in various topics such as personality [Kosinski et al. 2013], happiness [Dodds et al. 2011], and social influence [Lewis et al. 2012]. Our work provides an efficient approach for profession identification that will benefit profession-related research in sociology.

4.2. User Profiling

User profiling aims to infer various attributes of users from social media [Mislove et al. 2011]. These attributes can be roughly divided into explicit attributes (e.g., gender and age) and implicit attributes (e.g., interests, happiness and political orientation).

Existing user profiling studies mainly focus on explicit attributes and usually adopt classification and recommendation methods for attribute prediction. Most classification-based works devote to extract efficient features from UGC to predict specific attributes, such as gender and age [Goswami et al. 2009; Burger et al. 2011; Fink et al. 2012], location [Rao et al. 2010; Li et al. 2012], tags [Feng and Wang 2012; Liu et al. 2012; Tu et al. 2014], and other explicit labels [Kong et al. 2013; Chaudhari et al. 2014; Jacob et al. 2014].

Most explicit attributes are inferred from user-generated text data. For those attributes with rich sociality, social network structure may also be considered for prediction [Mislove et al. 2010; Yang et al. 2011; Li et al. 2012; Sachan et al. 2014; Tu et al. 2016a, 2016b]. Researchers are also interested in implicate attributes such as personal interests [Yang et al. 2011], political orientation [Rao et al. 2010], personality traits [Golbeck et al. 2011; Schwartz et al. 2013], and social power [Danescu-Niculescu-Mizil et al. 2012]. Yang et al. [2015] proposed to learn network representations by incorporating text information and network structure. However, it cannot handle multiple sources of text information.

This article focuses on profession identification with heterogenous text information, which has been less studied by previous work. Traditional multi-task learning approaches [Evgeniou and Pontil 2004, 2007] usually aim to learn a shared representation across multiple related tasks. In this work, we focus on a unique profession identification task and acquire the joint representations through a two-stage process.

As compared with existing work, our framework compressively considers personal information, community structure, and unlabeled data together to identify professions, which can be easily adapted to other social attributes of users.

5. CONCLUSION

This article presents an efficient framework PRISM for profession identification in social media. The proposed PRISM identifies professions with both personal information

and network structure and addresses several practical challenging issues including incorporating heterogeneous information and utilizing unlabeled data. The experiments on a large real-world dataset demonstrate the effectiveness of PRISM, which can be easily extended to identify other attributes. We also investigate characteristics of microblog users in various aspects including social networks and linguistic styles.

We plan to further explore the following research issues in the future:

(1) This article adopts a simple strategy, multi-training, to take advantages of unlabeled data. We will explore more sophisticated semi-supervised learning methods for profession identification.
(2) Multiple social attributes of users may interact with each other and exhibit complicated correlations. We will explore joint identification of personal attributes such as age, gender, locations, and professions.
(3) Profession, as an important social attribute of people, will significantly influence people's many aspects such as language usage. We will extensively investigate these effects and patterns, which will be of great significance for both sociology research and commercial services.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful and constructive comments.

REFERENCES

Demetris Antoniades, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos, Sotiris Ioannidis, Evangelos P. Markatos, and Thomas Karagiannis. 2011. We.b: The web of short URLs. In *Proc. WWW*. 715–724.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proc. EMNLP*. 1301–1309.

Janet Saltzman Chafetz. 1988. The gender division of labor and the reproduction of female disadvantage toward an integrated theory. *J. Family Iss.* 9, 1 (1988), 108–131.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3 (2011), 27.

Gaurish Chaudhari, Vashist Avadhanula, and Sunita Sarawagi. 2014. A few good predictions: Selective node labeling in a social network. In *Proc. WSDM*. 353–362.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proc. WWW*. 699–708.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE* 6, 12 (2011), e26752.

A. Evgeniou and Massimiliano Pontil. 2007. Multi-task feature learning. In *Proc. NIPS*, Vol. 19. 41.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proc. KDD*. 109–117.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* 9 (Aug. 2008), 1871–1874.

Wei Feng and Jianyong Wang. 2012. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *Proc. KDD*. 1276–1284.

Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *Proc. ICWSM*.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (March 2003), 1289–1305.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *Proc. CHI*. 253–262.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proc. ICWSM*.

John L. Holland. 1997. *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*. Psychological Assessment Resources.

Yann Jacob, Ludovic Denoyer, and Patrick Gallinari. 2014. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proc. WSDM*. 373–382.

Xiangnan Kong, Bokai Cao, and Philip S. Yu. 2013. Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *Proc. KDD*. 614–622.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15 (2013), 5802–5805.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and others. 2009. Life in the network: The coming age of computational social science. *Science* 323, 5915 (2009), 721.

Kevin Lewis, Marco Gonzalez, and Jason Kaufman. 2012. Social selection and peer influence in an online social network. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1 (2012), 68–72.

Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *Proc. KDD*. 1023–1031.

Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2012. Tag dispatch model with social network regularization for microblog user tag suggestion. In *Proc. COLING*.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.* 30 (2007), 457–500.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Ann. Rev. Sociol.* (2001), 415–444.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proc. ICWSM*.

Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. 2010. You are who you know: Inferring user profiles in online social networks. In *Proc. WSDM*. 251–260.

Mark E. J. Newman. 2006. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 23 (2006), 8577–8582.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *J. Lang. Soc. Psychol.* 21, 4 (2002), 337–360.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of Workshop on Search and Mining User-Generated Contents*. 37–44.

Robert A. Rothman. 1987. *Working: Sociological Perspectives*. Prentice-Hall Englewood Cliffs, NJ.

Mrinmaya Sachan, Avinava Dubey, Shashank Srivastava, Eric P. Xing, and Eduard Hovy. 2014. Spatial compactness meets topical consistency: Jointly modeling links and content for community detection. In *Proc. WSDM*. 503–512.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and others. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* 8, 9 (2013), e73791.

Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2014. Inferring correspondences from multiple sources for microblog user tags. In *Chinese National Conference on Social Media Processing*. Springer, 1–12.

Cunchao Tu, Hao Wang, Xiangkai Zeng, Zhiyuan Liu, and Maosong Sun. 2016a. Community-enhanced network representation learning for network analysis. *arXiv preprint arXiv:1611.06645* (2016).

Cunchao Tu, Weicheng Zhang, Zhiyuan Liu, and Maosong Sun. 2016b. Max-margin deepwalk: Discriminative learning of network representation. In *Proc. IJCAI*. 3889–3895.

Rudi Volti. 2011. *An Introduction to the Sociology of Work and Occupations*. Pine Forge Press.

Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y. Chang. 2015. Network representation learning with rich text information. In *Proc. IJCAI*. 2111–2117.

Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. 2011. Like like alike: Joint friendship and interest propagation in social networks. In *Proc. WWW*. 537–546.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proc. ICML*, Vol. 97. 412–420.

Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 3, 1 (2009), 1–130.

Received November 2015; revised December 2016; accepted March 2017