PRISM: Profession Identification in Social Media with Personal Information and Community Structure

Cunchao Tu, Zhiyuan Liu(✉), and Maosong Sun

State Key Lab on Intelligent Technology and Systems, National Lab for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
{tucunchao,lzy.thu}@gmail.com, sms@tsinghua.edu.cn

Abstract. User profession plays an important 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 paper, we explore the task of identifying user professions according to their behaviors in social media. The task confronts the following challenges which 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 of **PR**ofession **I**dentification in **S**ocial **M**edia (PRISM). It takes advantages 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 confidences 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 significant effectiveness of our method compared with other baseline methods.

1 Introduction

Social media services, such as microblogs, enable users to post messages, share information and communicate with each other in social networks. Besides, users may also contribute tags and short notes to describe themselves. The user generated content (UGC) reserves rich facts about users, including their 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 [2], to more complicated ones such as personality [23], happiness [6] and political polarity [21].

Profession, which is founded upon specialized educational training and aims to supply service to others, is also a critical social attribute of people. Sociologists

X. Zhang et al. (Eds.): SMP 2015, CCIS 568, pp. 15–27, 2015.
DOI: 10.1007/978-981-10-0080-5_2

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, self and social identity [24]. 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 recommendation 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. The task is the focus of this paper.

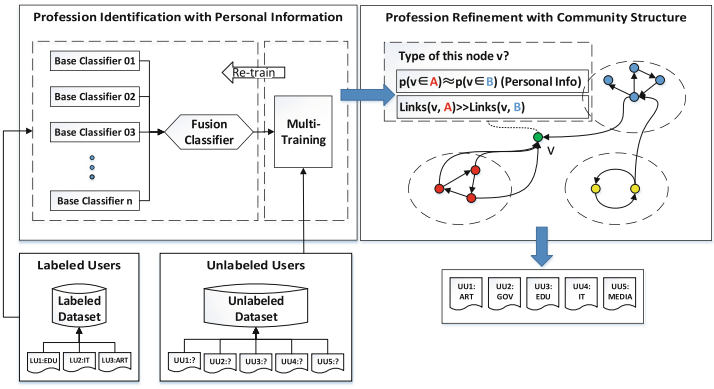

Fig. 1. The framework of PRISM.

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 paper, 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) *Personal Information.* Microblog users provide self descriptions and user tags, and constantly post short messages. The user generated contents form the personal information and can provide rich clues about user professions.
(2) *Network Information.* A user usually follows others to get information he/she is interested in. The following behaviors form social networks of microblog users. In our dataset we group users of the same professions into profession communities, which exhibit a relatively high modularity [20] score of 0.25. This indicates strong correlations between professions and network structure, and also confirms the homophily theory in sociology [17] that 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 these 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 of **PR**ofession **I**dentification in **S**ocial **M**edia (PRISM). PRISM takes advantages of both personal features and community structure, for user profession identification in social media.

Firstly, for heterogeneous personal information, we present a *cascaded two-level* classifier to measure confidences 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 final decision.Then, we further present a *multi-training* process, following the idea of co-training, to take advantages of both labeled and unlabeled users to improve classifier performance. Finally, we propose an profession identification method synthetically considering the confidences from personal features and community structure.

In the experiments, we collect more than 60 thousand 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" and "IT", etc. The experimental results on our dataset show that our method achieves the accuracy of 84.92 %, which outperforms all other baseline methods significantly.

2 The Framework of PRISM

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 advantages of profession community structure to refine profession identification. We introduce the details of our method as follows (Fig. 1).

2.1 Profession Identification with Personal Information

In this step, each user u is represented as a bag of feature vectors $\mathcal{X}_u = \{\mathbf{x}_{u,r}\}$. Here each $\mathbf{x}_{u,r}$ denotes a feature vector obtained from a distinct information source r, where $r \in \{1, \ldots, R\}$ and R is the number of information sources.

Suppose we have a set of annotated user-profession pairs $\{(\mathcal{X}_u, y_u)\}$ for training, where $y_u = k \in \{1, \ldots, K\}$ and K is the number of professions.

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 a 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.

Feature Design and Base Classifier Construction. In social media, a user generates various types of content. Taking the user "Kai-Fu Lee", a famous Chinese IT activist, for example, 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. These information should be handled separately due to their distinct characteristics. In this paper, we consider eight distinct sources of user generated personal information to build features for base classifiers, which are listed in Table 1.

Table 1. 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-word 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` which are usually in form of tiny URLs [1], 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 [10,26], we use χ^2 statistic to select representative features for each feature source. Afterwards, we build base linear classifiers for each feature source.

Base Classifier Fusion. The prediction result $\mathcal{P}_u$ obtained from base classifiers for user u is a matrix, which can not be directly used as input of fusion classifier. We concatenate the transfer matrix $\mathcal{P}_u$ into a feature vector as input of fusion classifier, i.e., building a feature vector $\mathbf{z}_u$ simply by concatenating column vectors of $\mathcal{P}_u$, i.e., $z_{u,k+K\times(r-1)} = p_{k,r}$. 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 the concatenation scheme significantly outperforms the other schemes (max and sum), hence we only report the concatenation results.

We select Liblinear [7][1] 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 SVM[2], and LR performs better in both effectiveness and efficiency. Hence, in the following part we only show the results obtained with LR.

Multi-training with Labeled and Unlabeled Data. In real world, there are 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, we use them to identify professions for unlabeled users. We select the users that more than *half* base classifiers agree on their professions, and put these users with corresponding identified profession labels into training set. Then we re-train these base classifiers.

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.

2.2 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

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

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

and sociology theory [17]. Following [19], we assume that users of the same profession form an profession-specific community. A relatively high modularity [20] 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, i.e., 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, i.e., *structure quality* and *content quality*.

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 $[-1, +1]$, and a strongly positive score indicates there is significant community structure in G_k.

This measure is originally proposed by [19] to compute the quality of a community, named as *normalized conductance*. In this paper, we integrate it with content quality together for profession refinement.

Content Quality. Content quality measures the significance of a community based on personal confidences of all users assigned in this community. In this paper, 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.

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 [19].

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 putted into an 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 professions of unlabeled users.

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 Weibo[3]. The ratios of various professions among these users are shown in Table 2. We also collect 150,000 verified users with no profession annotations for multi-training.

From the profession composition of these labeled users, we find that the users in "media" and "government" are dominant. The reasons may be: (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 online and contact with citizens officially. Therefore, many national and local officials have registered in Sina Weibo.

3.1 Experimental Results on Profession Identification

We randomly divide the 62,415 labeled users into 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

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

Table 2. 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

identification for U_{correct} users, 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, for profession k precision 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 $\frac{2PR}{P+R}$.

Profession Identification with Personal Information. 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, 6600 features for MSG, 3200 features for MEN, 2700 features for URL, 3600 features for ENT and 4100 features for HAS, which achieve the best performance for each base classifier.

Table 3 shows the evaluation results on profession identification with various features of their combinations. In this table, the line of "Single Vector" is the baseline which represents a user by taking 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 3, 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 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.
(3) The fusion classifier achieves much better performance compared to all base classifiers. This indicates that, the fusion of base classifiers can significantly improve identification.

(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.

Table 3. 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**

Profession Refinement with Community Structure. To evaluate the performance of our profession refinement with community structure, we take two community-based methods as baselines, i.e., label propagation algorithm (LPA) and community detection (CD). LPA addresses the task as graph-based semi-supervised learning [27]. The basic idea of LPA is the labels of a user is dependent on its neighbors. By propagating labels from annotated users to unannounced users through a social network, LPA can identify profession labels of users. As previously introduced, CD is the user profiling algorithm proposed in [19]. The both methods only consider community structure to classify users.

We show the evaluation results in Table 4. From the table we observe that:

(1) Profession refinement with community structure achieves considerable improvement as compared to the two-level classifier. This indicates the community structure can also provide supplementary information for profession identification beyond personal information.
(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.

Table 4. 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

	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.

To investigate the reason of identification errors, we show the distribution of identified professions for each profession in Fig. 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 profession j, i.e.,

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

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.

4 Related Work

User profiling aims to infer various attributes of users from social media [18]. 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 [2,9,12], location [15,21], tags [8,16] and other explicit labels [4,13,14].

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 [15,19,22,25]. Researchers are also interested in implicate attributes such as personal interests [25], political orientation [21], personality traits [11,23] and social power [5].

This paper focuses on profession identification, which has been less studied by previous work. As compared with existing work, our framework compressively consider 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 paper 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 plan to further explore the following research issues in the future: (1) This paper 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.

Acknowledgement. This work is supported by the National Natural Science Foundation of China under Grant Nos. 61170196 and 61202140 and the Major Project of the National Social Science Foundation of China under Grant No. 13&ZD190.

References

1. Antoniades, D., Polakis, I., Kontaxis, G., Athanasopoulos, E., Ioannidis, S., Markatos, E.P., Karagiannis, T.: we.b: the web of short URLs. In: Proceedings of WWW, pp. 715–724 (2011)
2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of EMNLP, pp. 1301–1309 (2011)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM TIST **2**(3), 27 (2011)
4. Chaudhari, G., Avadhanula, V., Sarawagi, S.: A few good predictions: selective node labeling in a social network. In: Proceedings of WSDM, pp. 353–362 (2014)
5. Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., Kleinberg, J.: Echoes of power: language effects and power differences in social interaction. In: Proceedings of WWW, pp. 699–708 (2012)
6. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. PLoS ONE **6**(12), e26752 (2011)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. JMLR **9**, 1871–1874 (2008)
8. Feng, W., Wang, J.: Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In: Proceedings of KDD, pp. 1276–1284 (2012)
9. Fink, C., Kopecky, J., Morawski, M.: Inferring gender from the content of tweets: a region specific example. In: Proceedings of ICWSM (2012)
10. Forman, G.: An extensive empirical study of feature selection metrics for text classification. JMLR **3**, 1289–1305 (2003)
11. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: Proceedings of CHI, pp. 253–262 (2011)
12. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Proceedings of ICWSM (2009)
13. Jacob, Y., Denoyer, L., Gallinari, P.: Learning latent representations of nodes for classifying in heterogeneous social networks. In: Proceedings WSDM, pp. 373–382 (2014)
14. Kong, X., Cao, B., Yu, P.S.: Multi-label classification by mining label and instance correlations from heterogeneous information networks. In: Proceedings of KDD, pp. 614–622 (2013)
15. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of KDD, pp. 1023–1031 (2012)
16. Liu, Z., Tu, C., Sun, M.: Tag dispatch model with social network regularization for microblog user tag suggestion. In: Proceedings of COLING (2012)
17. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Ann. Rev. Sociol. **27**, 415–444 (2001)
18. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the demographics of twitter users. In: Proceedings of ICWSM (2011)

19. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of WSDM, pp. 251–260 (2010)
20. Newman, M.E.: Modularity and community structure in networks. PNAS **103**(23), 8577–8582 (2006)
21. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of Workshop on Search and Mining User-Generated Contents, pp. 37–44 (2010)
22. Sachan, M., Dubey, A., Srivastava, S., Xing, E.P., Hovy, E.: Spatial compactness meets topical consistency: jointly modeling links and content for community detection. In: Proceedings of WSDM, pp. 503–512 (2014)
23. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE **8**(9), e73791 (2013)
24. Volti, R.: An Introduction to the Sociology of Work and Occupations. Pine Forge Press, Thousand Oaks (2011)
25. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of WWW, pp. 537–546 (2011)
26. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. Proc. ICML **97**, 412–420 (1997)
27. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. **3**(1), 1–130 (2009)

http://www.springer.com/978-981-10-0079-9

Social Media Processing
4th National Conference, SMP 2015, Guangzhou, China, November 16-17, 2015, Proceedings
Zhang, X.; Sun, M.; Wang, Z.; Huang, X. (Eds.)
2015, XIV, 242 p. 76 illus. in color., Softcover
ISBN: 978-981-10-0079-9