Proceedings of the 3rd International Workshop on Social Influence Analysis (SocInf 2017)
August 19th, 2017 - Melbourne, Australia

Identifying Influential Users' Professions via the Microblogs They Forward

Yuan Wang, Hangyu Mao, and Zhen Xiao

Department of Computer Science, Peking University, Beijing 100871, China
{wangyuan, mhy, xiaozhen}@net.pku.edu.cn

Abstract. For most social media sites, how to find out (influential) users' professions is an important task. Much work has been conducted to explore this task through mining user-generated textual content or analyzing the social network structure. In this paper, we innovatively solve this task by only examining which microblog messages an influential user has forwarded. First, we define hot microblog messages under two standards and identify them from a large number of candidate messages. Each of the identified messages points to a specific hot event. Next, we group similar hot messages together based on their word similarity, semantic similarity, and forwarders' similarity. Last, we represent users with the hot messages they forwarded and design an identification method to identify their professions. Moreover, we collect a real-world dataset to conduct experiments and prove that our method performs significantly better than the traditional method.

1 Introduction

Online microblogging services have become an integral part of the daily life for most Netizens. These services expect to know more about their users' profiles, since user profile plays an important role in commercial services, such as personalized recommendation and online advertising. However, user profile is usually not easily obtained, because users are reluctant to expose their profiles to the public. Fortunately, some work has been conducted to solve this problem. A traditional practice is cutting users' messages into bags of words and training a classifier. This practice can achieve an acceptable result on simple tasks such as predicting gender and age [1], but it can not solve more complex tasks [14].

Profession, which is founded upon specialized educational training, is a critical social profile of influential users. In Weibo, the largest microblogging service in China, influential users are mainly organized by their professions. They are more likely to follow other users that have the same profession with them. It is important to correctly identify influential users and their professions for microblogging services.

Message forwarding (e.g. retweeting on *Twitter.com* and reposting on *Weibo.com*) is one of the most popular functions in the existing microblogging services. In Weibo, users can forward messages or any interesting content on the web, such as real blogs, photos and external links. In this paper, if a weibo message was forwarded by any user, we define it as forwarded message, otherwise we define it as non-forwarded message. Based on a large dataset, we find that about 60% of weibo messages are forwarded messages. For most users, the messages they forwarded are exactly what they are interested in. Users' professions can be reflected by the messages they forwarded to some extent. But the traditional "bag of words" model will completely undermine the information contained in users' forwarding behaviors. Naturally, in this paper, we ask and try to answer the following question: can we represent microblog users with the messages they forwarded, and predict their professions more accurately than the traditional method?

The task confronts some challenges which make it non-trivial. The first challenge is that there exist too many forwarded messages. If we consider each forwarded message as a feature, the feature vector will be very large and sparse. We observe that most of these messages only have been forwarded by no more than 3 weibo users. In this paper, we define them as non-hot forwarded messages and define other messages that are forwarded by more users as hot forwarded messages. In our experiment, we discard the non-hot messages. Another challenge is that even though we can filter out non-hot messages, the number of remaining hot messages is still quite large. We observe that, every hot message points to a hot event (e.g. a breaking news or a recently released movie). We should come up with some methods to group similar hot weibo messages together.

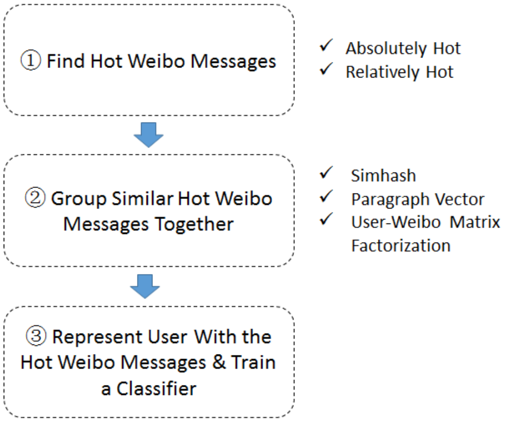

① Find Hot Weibo Messages
✓ Absolutely Hot
✓ Relatively Hot
② Group Similar Hot Weibo Messages Together
✓ Simhash
✓ Paragraph Vector
✓ User-Weibo Matrix Factorization
③ Represent User With the Hot Weibo Messages & Train a Classifier

Fig. 1. The framework of PIFB

In this paper, we propose an efficient framework of Profession Identification by using Forwarding Behaviors (PIFB). As Figure 1 shows, first, we identify the hot forwarded messages from a large number of candidates. Each of these identified messages points to a specific hot event. Next, we introduce three methods to group similar messages together, downsizing our message sets. Then, influential users can be represented with the merged hot messages that they have forwarded. Finally, we predict users' professions, and the results are more accurate than those in the traditional method.

2 Dataset and Professions

We collect 41,531 manually annotated influential users from Weibo (http://weibo.com). To avoid robot users, we only collected verified users. Weibo conducts manual verifications to make sure that the verified users provide real and authentic information. These users belong to 11 representative professions. As Table 1 shows, the professions include "media", "entertainment", "sports", and "IT", etc.

We also collect users' latest 500 weibo messages. These messages can be classified into two categories: forward action and post action. In general, forward action consists of *trace* and *content*. Trace contains the information that through which users the current user can see the final messages. Content can be extended to any forms as long as it can be shared by users with their followers, such as videos and blogs. A simple example is shown below: if a user froward the message:

$$\underbrace{RT\ \ @Raj\ \ RT\ \ @Sheldon}_{trace} : \underbrace{It\ \ took\ \ 50\ \ years\ \ ...}_{content}$$

This forward action indicates that "It took 50 years ..." was originally posted by "Sheldon" and was forwarded by "Raj", and now is forwarded by the current user again.

In general, post action only contains the "content" part, representing that the current user posted an original message.

3 The Framework of PIFB

In this section, we formalize our problem as a classification task and introduce the main steps of PIFB.

3.1 Hot Message Identification

This paper focus on influential user's behaviors about the forwarded messages. A critical step is to identify the hot forwarded messages. In this part, we define hot messages under two standards.

Absolutely Hot Message We argue that if a message has been forwarded by more users, the information behind it will be more. And the forwarding behaviors about this message can help our profession prediction more. Nowadays, Weibo has become the the biggest "News Site" in China. Most traditional news organizations open their official accounts in Weibo and these accounts are all very active. They usually publish the breaking news timely and make the news spread quickly. There also exist many Chinese celebrities in Weibo, including actors, singers and entrepreneurs, etc. They post their personal views or daily lives in their accounts. They generally have a great number of followers and their daily updates are likely to get thousands of forwards. So, in this paper, if a weibo message has been forwarded by more than a certain times (for example, 500), it will be regarded as the first kind of "hot forwarded message" (absolutely hot).

Table 1. The distribution of professions in our dataset.

No.	Category	(%)	No.	Category	(%)
1	Media	26.3	7	Sports	6.4
2	Entertainment	10.1	8	Fashion	6.2
3	Estate	9.1	9	Education	5.9
4	Finance	8.6	10	Literature	5.4
5	Government	8.5	11	Game	5.1
6	IT	8.4			

Relatively Hot Message The 11 professions, showed in Table 1, are not "evenly matched" on attracting attentions. Nearly all the high forwarded messages are all posted by "entertainment" and "sports" stars. For an "estate" account, it is not easy to post an absolutely hot message, because "estate" accounts usually have relatively less followers and lower forwarding rate. If we only adopt the absolutely hot messages as described in the previous paragraph, it is very possible that we only get the messages posted by a small subset of that 11 categories (may be 2-4). Therefore, as a supplement to the first standard, we define another kind of hot message. In our dataset, if a message's owner has f followers (f>500) and this message has been forwarded by more than $f/5$ times, it will be regarded as the second kind of "hot forwarded message" (relatively hot).

After identifying all these two types of "hot messages", we can build a matrix M, whose columns denote hot messages and rows denote users. This matrix represents all the forwarding relationships between weibo users and hot messages. M will have too much columns, if we don't filter out the non-hot messages. Even though we do only consider the hot messages, the number of column is also very big. To slim down M, we propose three methods to group similar messages together in the next.

3.2 Group Similar Hot Messages Together

In most microblogging services, users can be divided into two categories: information producer and information consumer. The information producer mainly includes the news site accounts, self-media accounts, and profit-seeking accounts with legions of followers. Their main purpose is making their microblogs broadcast as widely as possible to expand their influence and get more new followers. Whenever there is a news, producers will timely post their relevant microblogs. The producers are very likely to post similar contents, because the texts may be pasted from the same source. The information consumer mainly refers to normal weibo users. More than 90% weibo users can be classified into this category. Their most important action is reading and forwarding messages. Normally, hot messages are more likely to attract them.

If the hot messages only contain a video link or a web link, it is easy to determine whether they are similar. But if they contain some text contents, the task will be more difficult. In the next, we introduce three methods to solve it.

Simhash As described above, the information producers are likely to post similar weibo messages. The most direct idea is that merging similar hot messages based on their word similarity. Simhash [2] is a widely used dimensionality reduction technique in calculating the document similarity. This model can map high dimensional document vectors to small-sized fingerprints. With the help of simhash, we can transform such a high-dimensional vector into a k-bit fingerprint where k is quite small, such as 64. An important characteristic of simhash is that, similar documents have similar hash values. For instance, if there are two documents that only differ in a single word, the cryptographic hash functions will hash them into two completely different values. However, simhash will hash them into similar fingerprints. This characteristic is very important in calculating the document similarity.

In this method, we firstly calculate the simhash values of all the hot messages. After that, we can group the similar messages together, if the hamming distance of their simhash fingerprints is less than or equal to 3.

Paragraph Vector The simhash can only calculate the documents' similarity based on their word similarity. It can not deal with situation that, two documents have the similar semantics but written with different words. [8] proposes "Paragraph Vector" (P2V), an unsupervised framework that learns continuous distributed vector representations for pieces of texts. This method can be applied to variable-length paragraphs, and transform them into fixed-length vectors. In this model, every weibo message is mapped to a unique vector, represented by a column in a matrix and every word is also mapped to a unique vector, represented by a column in another matrix. The paragraph vectors and word vectors are concatenated to predict the next word. They are trained using s-tochastic gradient descent and the gradient is obtained via backpropagation. Details can be found in the original paper. After being trained, the distance between two paragraph vectors will be small if they talk about a same topic. It is not sensitive about the synonym. These vectors can be used as features directly to conventional machine learning models, such as logistic regression or k-means.

We firstly calculate hot messages' representative vectors by using the "Paragraph Vector" method. The length of vector is set to 400 according to the original paper. After that, we calculate their distances. A pair of hot messages can be grouped together if their distance is smaller than a threshold.

User-Weibo Matrix Factorization The first method is based on message's word similarity and the second is based on the semantic similarity. They are both directly calculated by the weibo contents. As described in section 3.1, we have generated the user-weibo relationship matrix M. So we can further find more similar messages based on which users have forwarded these messages. Hofmann [5] introduced the PLSA, which developed probabilistic latent semantic models for performing collaborative filtering. In this step, PLSA models users ($u{\in}U$) and documents ($d{\in}D$) as random variables, taking values from the space of all possible users and documents respectively. The relationship between them is learned by modeling the joint distribution of users and documents as a mixture distribution. The hidden variables t ($t{\in}T$, $\|T\|{=}k$) represent the topics between U and D. The model can be written in the form of mixture model as the next equation:

$$P(u|d;\theta) = \sum_{t=1}^{k} p(u|t)p(t|d) \tag{1}$$

Based on this model, we can transform the user-weibo matrix into two new matrices. The first is user-topic matrix, which represents each user with a vector of k topics. The second is document-topic matrix, which represents each document with a vector of k topics too. In the second matrix, if the documents contain similar topics, their vectors are more likely similar. We can group two similar hot messages together, if the distance between their vectors is under a threshold. In this paper, we empirically set k to 400 and name this method UWMF.

3.3 Profession Prediction

After merging similar hot messages, users can be represented as more compact vectors. Each element of these vectors represents a merged hot message, and the elements will be used as features in our multi-class classifier.

Over the last several decades, many kinds of discriminant classifier have been created. In our experiment, we compare Logistic Regression (LR) and Gradient Boosted Decision Tree (GBDT). We choose GBDT as our default multi-class classifier, because we find that GBDT performs better in most instances. Hence, in the following part we only show the results obtained with GBDT [3].

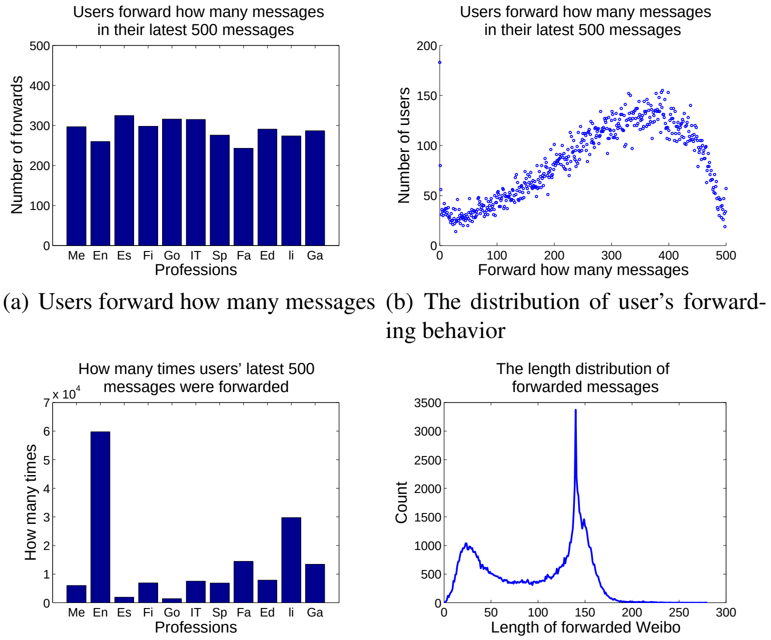
4 Experiment Results

In this section, we first statistically study our dataset. After that, we identify the hot weibo messages and merge the similar ones. At last, we compare our methods with the baseline method comprehensively.

4.1 Observation

We firstly count influential user's forwarding rates on different professions. As Figure 2(a) shows, different professions have different forwarding rates on average. It is a little surprise that the “estate” and “government” accounts forwarded more messages compared with the “finance” accounts. Overall, the difference between different professions is not significant. In our dataset, about 58% of weibos are all forwarded messages. For about 66% users, more than half of their messages are forwarded messages. Figure 2(b) shows the distribution of how many messages users forwarded (in their latest 500 messages) in our dataset. We find that about 95% users forwarded more than 50 messages. In this paper, our goal is to predict users professions only based on their forwarding behaviors, so we discard other 5% users who forwarded no more than 50 messages in our experiment.

As described in section 3.1, we define the absolutely hot message and the relatively hot message separately. To better understand these two types, we calculate how many times that users' latest 500 weibo messages have been forwarded on average by category. As Figure 2(c) shows, these numbers of different categories are very unbalanced. The “entertainment” and “literature” accounts attract much more forwarding behaviors than "estate" accounts. The main reason is that the "entertainment" and "literature" accounts have relatively more followers. If we only adopt absolutely hot messages (for example, the threshold is set to 500), it is possible that we can not get any hot messages posted by "estate". So identifying relatively hot messages is very necessary in our model.



(a) Users forward how many messages (b) The distribution of user's forwarding behavior

(c) Number of users' 500 messages were forwarded (d) Length distribution of forwarded messages

Fig. 2. Data observation

Weibo limits message length to 140 Chinese characters or 280 English characters. Figure 2(d) shows the length distribution of hot messages in our dataset. We can find that there exist two peaks. The first peak represents the hot messages that only contain 10-20 characters. These messages are likely to be posted by star users who have millions of fans. This kind of message usually additional contains a picture or a video link. The second peak represents the messages that contain 140 Chinese characters. This kind of message generally contains rich semantics.

4.2 Identify Hot Messages

As described in section 3.1, if a message has been forwarded by more than a certain number of times, it will be considered as an absolutely hot message. It is apparent that how to set the threshold is a double-edged sword. If we set the threshold to a smaller value (more hot messages), on one hand, user can be represented with more messages and our model's expression ability will be increased; on the other hand, our model should handle more features and need to take the risk of over-fitting. As Table 2 shows, we set the threshold to 500, 2,000, and 10,000 separately. When the threshold is set to 500, we can get 731,153 hot weibo messages. This number is too large and most of these messages have been forwarded by no more than 5 users in our dataset (40 thousand users). Then, we filter out such messages from our hot message sets, leaving 100,219 valid messages. In the prediction tasks, we compare the performance of these three thresholds and choose 500 as the default value.

Table 2. Number of absolutely hot messages

No.	Threshold	# before filter	# after filter
1	500	731,150	100,219
2	2000	426,019	82,339
3	10000	74,308	32,955

As section 3.1 shows, if a message's owner has f followers (f>500) and this message has been forwarded by more than $f/5$ times, we regard this weibo message as a relatively hot message. Just as the absolutely hot messages, we also filter out the messages that have been forwarded by no more than 5 users in our dataset, and get 61,806 relatively hot messages.

Eventually, we collect 162,025 hot messages in total (100,219 absolutely hot & 61,806 relatively hot).

4.3 Group Similar Hot Messages Together

In this part, we evaluate the performance of our three methods on clustering similar hot messages. As Table 3 shows: (1) In the simhash method, we choose 64 as the default length of hash value. In this step, we group similar messages together, if their hamming distance is less than or equal to 3. We can merge our 162,025 hot messages, identified from section 4.2, into 57,624 hot events. (2) In the second method, we choose 400 as the default size of paragraph vector, and merge similar messages according to their Euclidean distances. In this step, we can merge the 162,025 hot messages into 32,118 hot events. (3) In the third method, we also choose 400 as the size of hidden variables, and adopt Euclidean distance to measure their similarities. In this step, we can merge the 162,025 hot messages into 27,129 hot events. In our experiment, the lengths of these three vectors (64, 400, 400) are chosen empirically [8, 10]. We validate the other hyper-parameters (where to stop merging) with the validation set, and find the best stop points.

In practice, we serially combine all these three methods. At first, we adopt the simhash to find similar hot messages, making users' representative vectors more compact. On the basis of this results, we adopt the second method, further compressing users' vectors. At last, we perform the third method based on the current results. After

Table 3. Number of messages under different merging strategies

No.	**Merging Strategy**	**# before**	**# after**
1	Simhash	162,025	57,624
2	P2V	162,025	32,118
3	UWMF	162,025	27,129
4	Simhash+P2V+UWMF	162,025	17,196

these three steps, our 162,025 hot messages can cluster together into 17,196 hot events. In the next, we will study whether these optimizations can improve our profession identification tasks.

4.4 Results of Prediction

We randomly divide our 40 thousand labeled users into training set (60%), validation set (20%), and test set (20%). We regard user's labeled profession as the gold standard, and select accuracy, macro-averaging precision/recall/F-Measure as evaluation metrics.

To verify the validity of our method, we build a baseline model. The feature candidates of baseline model include: (1) Words in user's original messages; (2) Words in user's forwarded messages; (3) Mentioned user ids in messages; (4) URLs in messages; (5) Hash tags in messages. There exist hundreds of thousands of feature candidates and we have to perform feature selection to downsize our feature sets. Following the valid experience in feature selection for text classification, we use χ^2 statistic to select representative features. We evaluate performance with different numbers of features, and select 9200 feature candidates. We compare LR and GBDT on these features and find they have similar performance. To be consistent with our model, we also choose GBDT as the default baseline classifier.

Table 4. Evaluation results for various features and combinations. (%)

No.	**Method**	**Accuracy**	**Precision**	**Recall**	**F1**
1	Baseline	62.38	64.03	60.29	62.10
2	Simhash	69.24 ↑ 6.86%	70.88	67.61	69.21 ↑ 7.11%
3	Simhash+P2V	73.79 ↑ 11.41%	73.90	71.28	72.57 ↑ 10.47%
4	Simhash+P2V+UWMF	73.98 ↑ 11.60%	74.81	72.95	73.87 ↑ 11.77%

From Table 4, we can observe the evaluation results. We find that the baseline model achieves a performance of 62.38% in accuracy and our three models all get better results than it. This comparison proves user's forward behavior is effective in profession identification. As Table 4 shows, along with the implementation of three merging strategies, our three models can make the prediction gradually improved. Our model in the fourth line that serially adopts all three merging strategies achieves the best result (accuracy=73.98, F1=73.87). This result indicates that effective clustering of similar messages is necessary, for there exist too many forwarded messages.

To better understand the prediction errors, we present the details of the best result. In Table 5, the value of i_{th} row and j_{th} column represents the ratio of the users in profession i being identified as profession j.

Table 5. Distribution of identified professions in each profession.

	Me	**En**	**Es**	**Fi**	**Go**	**IT**	**Sp**	**Fa**	**Ed**	**Li**	**Ga**
Me	76.7	5.6	2.7	3.5	4.1	3.3	2.2	0.9	0.6	0.2	0.2
En	7.2	74.5	0.2	3.3	0.7	1.4	4.4	5.1	0.2	1.3	1.7
Es	7.4	2.0	72.9	8.5	5.3	2.2	0.4	0.9	0.1	0.0	0.3
Fi	8.4	0.1	6.4	70.2	5.3	6.2	0.2	1.3	1.7	0.1	0.1
Go	4.9	2.2	0.4	4.2	78.2	2.9	4.1	0.4	2.5	0.2	0.0
IT	6.1	0.7	3.9	4.3	1.3	76.3	0.2	0.1	2.6	0.7	3.8
Sp	5.1	2.9	0.0	0.3	0.3	1.0	86.2	2.2	0.7	0.0	1.3
Fa	9.7	14.9	1.0	6.2	0.2	0.0	3.3	61.5	0.9	1.2	1.1
Ed	5.2	3.9	3.3	4.6	2.0	3.2	0.7	1.8	68.4	4.2	2.7
Li	13.7	7.2	0.7	1.3	0.6	1.4	0.4	3.3	9.8	60.9	0.7
Ga	5.2	3.9	0.8	0.0	0.4	7.1	1.2	4.3	0.1	0.3	76.7

To make the data more intuitive, we illustrate the ratio in each entry using different shades of color. We can observe that: (1) Our model performs differently on different professions. The recall scores (value on the diagonal) of most professions are bigger than 70%, with only "fashion" and "literature" less than 65%. The main reason is that the forward behavior of these two professions has no special characteristics. (2) The "media" accounts occupy about a quarter of our user collections. Our model tends to predict the uncertain user as "media" account, making the precision score of "media" relatively lower (51.3%). (3) The behaviors of some professions are quite similar. For example, the "entertainment" user and "fashion" user have the similar interests, they usually follow and interact with each other. It makes the boundary between these two professions not very clear for identification.

5 Related work

User's attributes can be inferred from user-generated text data and social network structure. [6] showed that users' age and gender can be predicted from people's webpage browsing logs. [9] showed users' profiles can be predicted by their mobile phone apps. [13] analyzed tens of thousands of blogs and indicated significant differences in writing style and word usage between different gender and age groups. [1, 11] predicted user's gender and age based on their twitter linguistic characteristics. [15] identified weibo users' profiles only via the videos they talk about. [12] identified users' political orientation and ethnicity by leveraging their network structure and linguistic characteristics. [4, 17] predicted users' profiles based on their social network structure and chick ins.

Recently, there are some researches on identify users' professions. [14] presented an efficient framework for profession identification in Weibo. This work identified users' professions based on both personal information and network structure. [7, 16] showed that computers' judgments of people's personalities based on their Facebook Likes are more accurate than judgments made by their close acquaintances.

6 Conclusion

In this paper, we present an efficient framework PIFB to predict influential users' professions by only examining which microblogs they have forwarded. In the first step, we identify the hot weibo messages from a large number of candidate messages, and represent users with the hot messages they forwarded. After that, we group hot messages together if they talk about the similar topics. This step can make users' representative vectors more compact. At last, we design a multi-class classifiler to predict their professions. The experiments on a real-world dataset demonstrate the effectiveness of PIFB. Our method performs significantly better than the traditional "bag of words" based method.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work was supported by the National Natural Science Foundation of China under Grant No.61572044. The contact author is Zhen Xiao.

References

1. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the EMNLP. pp. 1301–1309 (2011)
2. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing. pp. 380–388. ACM (2002)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of SIGKDD. pp. 785–794. ACM (2016)
4. Culotta, A., Kumar, N.R., Cutler, J.: Predicting the demographics of twitter users from website traffic data. In: Proceedings of AAAI. pp. 72–78 (2015)
5. Hofmann, T.: Latent semantic models for collaborative filtering. ACM Transactions on Information Systems (TOIS) 22(1), 89–115 (2004)
6. Hu, J., Zeng, H.J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behavior. In: Proceedings of WWW. pp. 151–160. ACM (2007)
7. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences 110(15), 5802–5805 (2013)
8. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of ICML. pp. 1188–1196 (2014)
9. Malmi, E., Weber, I.: You are what apps you use: Demographic prediction based on user's apps. arXiv preprint arXiv:1603.00059 (2016)
10. Manku, G.S., Jain, A., Das Sarma, A.: Detecting near-duplicates for web crawling. In: Proceedings of WWW. pp. 141–150. ACM (2007)

11. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: ”how old do you think i am?”; a study of language and age in twitter. In: Proceedings of ICWSM. AAAI Press (2013)
12. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. In: Proceedings of ICWSM. pp. 281–288 (2011)
13. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
14. Tu, C., Liu, Z., Sun, M.: Social Media Processing: 4th National Conference, SMP 2015, Guangzhou, China, November 16-17, 2015, Proceedings, chap. PRISM: Profession Identification in Social Media with Personal Information and Community Structure, pp. 15–27. Springer Singapore, Singapore (2015)
15. Wang, Y., Xiao, Y., Ma, C., Xiao, Z.: Improving users’ demographic prediction via the videos they talk about. In: Proceedings of EMNLP (2016)
16. Wu, Y., Kosinski, M., Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences 112(4), 1036–1040 (2015)
17. Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X.: You are where you go: Inferring demographic attributes from location check-ins. In: Proceedings of WSDM. pp. 295–304. ACM (2015)