

Age Detection for Chinese Users in Weibo

Li Chen, Tieyun Qian^(✉), Fei Wang, Zhenni You,
Qingxi Peng, and Ming Zhong

State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China
{whulichen, feiw14, whuznyou}@163.com, {qiy, clock}@whu.edu.cn,
pengqingxi@gmail.com

Abstract. Age is one of the most important attributes in one user's profile. Age detection has many applications like personalized search, targeted advertisement and recommendation. Current research has uncovered the relationship between the use of western language and social identities to some extents. However, the age detection problem for Chinese users is so far unexplored. Due to the cultural and societal difference, some well known features in English may not be applicable to the Chinese users. For example, while the frequency of capitalized letter in English has proved to be a good feature, Chinese users do not have such patterns. Moreover, Chinese has its own characteristics such as rich emoticons, complex syntax and unique lexicon structures. Hence age detection for Chinese users is a new big challenge.

In this paper, we present our age detection study on a corpus of microblogs from 3200 users in Sina Weibo. We construct three types of Chinese language patterns, including stylistic, lexical, and syntactic features, and then investigate their effects on age prediction. We find a number of interesting language patterns: (1) there is a significant topic divergence among Chinese people in various age groups, (2) the young people are open and easy to accept new slangs from the internet or foreign languages, and (3) the young adult people exhibit distinguished syntactic structures from all other people. Our best result reaches an accuracy of 88% when classifying users into four age groups.

Keywords: Age detection · Chinese users · Feature selection · Feature combination

1 Introduction

In recent years, there is an increasingly research interest in user profiling in social media [4, 7–9, 11–13, 16]. This task is about predicting a user's demographics from his/her writing, mainly including gender, age, occupation, education, and political orientation. Among which, gender and age are two natural attributes and attract the most research attention. Compared to the problem of gender classification [1–3, 10, 14, 20], age detection is much more difficult and less examined. One reason is that it is hard to collect the labeled age data. As the age information is more sensitive and personal, many users tend to hide it in their profile.

The other reason is the lack of distinct features for age classification. In contrast, there are a bunch of gender features. For example, female users are generally more emotional than male users, and thus many sentimental words can be used to as the identifier of gender. Both these result in less studies on detecting age than on gender. Several pioneering work aims to reveal the relationship between users' age and the language use [4, 5, 11–13, 16–19]. However, existing researches are all based on western languages, i.e., English and Dutch. The age detection problem for Chinese users is so far unexplored.

Due to the cultural and societal difference, some well known features in English may not be applicable to the Chinese users. For example, while the frequency of capitalized letter in English has proven to be a good feature, Chinese users do not have such patterns. More importantly, Chinese has its own characteristics such as rich stylistic expressions, complex syntax and unique lexicon structures. There are about 161 western emoticons. In contrast, there are more than 2900 emoticons commonly used in the main social media in China. In addition, the document written in Chinese needs word segmentation to perform lexical analysis. Considering the fact that there are a number of informal use of language expression, the accuracy of word segmentation will be lowered down and this may affect the performance of word based model. Finally, the syntax structure for Chinese is also quite different from that for English. Will all these points lead to new challenges or chances for the age prediction task for Chinese users? What kind of features is of the most importance to this problem? To what extent can we identify a person's age group given his/her records in social media?

In this paper, we present our study to classify people into age categories. Since this is the first attempt to detecting ages of Chinese users, we have to build our own labeled data. For this purpose, we collect and annotate a corpus of microblogs from 3200 users in Sina Weibo, which is one of the biggest social networking sites in China. We then treat this as a supervised classification problem and derive models in three ways: 1) extracting stylistic features including emoticon, punctuation, and acronym; 2) using lexical based unigram features; and 3) using syntactic part-of-speech (POS) unigram structures. The word and POS unigram features are investigated to examine their effectiveness in Chinese in spirit of a fair comparison with their counterpart in English and Dutch [11–13, 15]. It should be also noted that the stylistic features used in [17], [4], and [5] are different from those used here in that they are treated as tokens as word unigrams rather than only a total occurrence, given the fact that we extract a large number of novel stylistic features from Chinese microblogs.

The contributions of this paper are as follows:

1. We present a first-of-a-kind age detection problem in Chinese social media. We collect and carefully build an annotated data set. We will publish this data set later for research use.
2. We construct a set of dictionaries, i.e., the stylistic, lexicon, and syntactic feature list. Then we systematically study their effects on age detection.

3. We find a significant topic divergence among Chinese people in various age groups. Furthermore, while old people tend to use conventional expressions, young and young adult people exhibit unique stylistic and syntactic structures, respectively.

Our study will provide new insights into the relationship between Chinese language and its users among different ages. It also has a great number of potential applications such as personalized search, targeted advertisement and recommendation.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the overall framework. Section 4 introduces three types of features. Section 5 provides experimental results. Section 6 analyzes some age-related interesting language patterns. Finally, Section 7 concludes the paper.

2 Related Work

With the rise of social media, user profiling has received considerable attention in recent years. There are a number of user identities in one user’s profile like gender, age, occupation, education, and political orientation. The overall framework for these tasks are all similar. We will review the literature in age detection in this section, organized by the age group, the feature set, and the classification method.

2.1 Age Group

Age prediction in most of existing studies is approached as a binary or multi-class classification problem. Rao et al. classify users into two major demographic pools, users who are below or above 30 [16]. Rosenthal and McKeown experimented with varying the binary split in order to find Pre- and Post-Social media generations [17]. With the research objective of detecting adults posing as adolescents, Peersman et al. set up the binary classification experiment for 16- vs. 16+, 16- vs. 18+, and 16- vs. 25+ [15]. When being posed as a multi-class problem, persons are categorized according to their life stages. Some of them are related to generations [4, 19], i.e., 10s, 20s, 30s, 40s, and higher, or multiple generations [11, 13], i.e., 20-, 20-40, and 40+. Others are based on life events (social age) and physical maturity (biological age). For example, Kabbur et al. use a biological age: kid (3..12 years), teen (13..17 years), young-adult (18..34 years), adult (35..49 years) and old (50+ years) [6]. Rosenthal and McKeown also use a three social age spans [17], namely, 18..22, 28..32, and 38..42 in their experiment. More recently, researches also start to predict age as a continuous variable and predicting lifestages [11–13].

In general, there are a number of ways to categorize people into age groups. Nevertheless, it is still problematic in finding clear boundaries. We adopt the three age groups in [17] and add one group for old people, which is close to the social age distribution in China.

2.2 Feature Set

Finding good features has long been a major interest for age detection. The content based features have been proved to be useful for detecting the age of users. The widely used ones are social-linguistic features [12, 16], character n-gram [15, 19], word n-gram [11–13, 15, 17, 19], pos n-gram [12], function words [5], and various stylistic features such as the number of punctuation and exclamation marks, slang words, or sentence length [4, 5, 17]. Besides the above features extracted from the texts, other features may also be used depending on whether their data sources contain such information. For example, the webpage structure is employed in [6], and the network structure and communication behavior are used in [16, 17].

While most of the studies are interested in comparing the effects of individual feature set, a few works pay attention to the combination of features [17]. We also note that feature selection is less examined in author profiling unlike that in text classification. The reason can be due to that the removal of common features may incur information loss in stylistic features. One exception is that in [15] which used χ^2 to do feature selection. Their results show that the best performance is achieved by using the largest informative feature set.

We extract the stylistic, lexical and syntactic features from Sina weibo, which are specific to Chinese language. We do not use structure and behavior features since we find they perform very poor in our preliminary experiment. Besides, we do feature selection and feature combination to evaluate their impacts.

2.3 Classification Method

A number of machine learning approaches have been explored to solve the age grading problem, including support vector machine (SVM) [15, 16], Naïve Bayes [4], logistic regression [11, 17]. When age is treated as a continuous variable, linear regression is adopted to train the model [1, 11, 12]. There are also a few literatures that study to combining the classifier built from different feature sets. Gressel et al. used a random forest classifier to ensemble learning approaches [5]. Rao et al. employed a stacked model to do simple classifier stacking [16].

The classification model is not the focus of this research. In our study, we choose to use SVM as our classifier.

3 Overall Framework

We follow a supervised learning framework for age detection. We look at four age groups containing a 5-year gap, i.e., young (18..22), young adult(28..32), adult(38..42), old(48..52). Each group is a category or class. A model is trained using features that are extracted from the contents of training users with annotated age spans, and then the model is used to predict the age group of test users. The overall framework is shown in Figure 1.

Since we are the first one studying this problem, there is no available data for this use. It takes us great efforts in building the corpus. Our data set consists of weibo downloaded from Sina, which is the largest social media in China. We

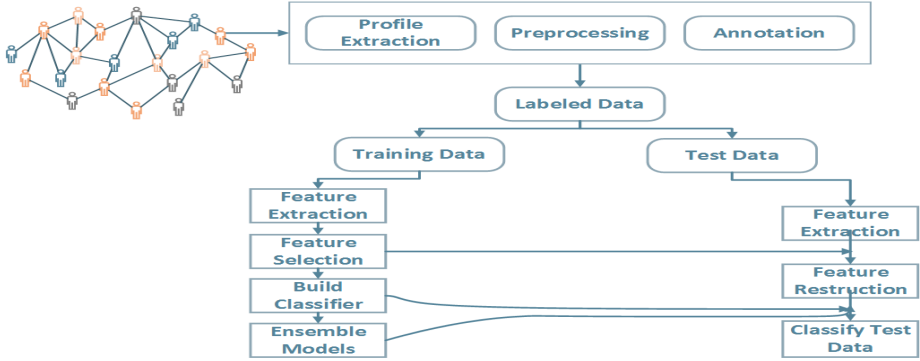


Fig. 1. The overall framework for age detection

use weibo as our testbed since the website provides the users the opportunity to post their age on their profile. We take advantage of this feature by downloading weibo where the user chooses to make this metadata publicly available. We downloaded 319,542 weibos containing age. However, the raw data is extremely imbalanced and contains a large number of noises. The dominant weibo users are those born between 1980 and 1990. There are more than 240,000 people in this span and less than 2,000 people aged over 50-year-old. So we randomly sample 1000 users for each group. We then place the following restrictions to pre-process the corpus:

1. The age in the profile should be valid (in an interval of $[10,80]$).
2. The account should represent an actual person, e.g. not an organization.
3. The account should have followers less than 1,000 to ensure that he/she is not a celebrity.
4. The account should have sufficient weibos (at least 10).

After the above filtering and pre-processing procedure, we then employ two students to manually check on a sample set of 300 profiles. The age labels are found to be correct for over 95% of them. Finally, we construct a corpus with 3200 users in 4 age classes, each having 800 users. Note that one user's weibos are merged into one document and then it is treated as an instance in each class.

4 Features

We extracted three types of features from users' weibos. They are treated as pseudo-words and their frequencies are counted to form various feature vectors for one user in terms of tf.idf values.

4.1 Stylistic Features

The stylistic features are extracted from the weibo. We use three types of stylistic features including emoticon, punctuation, and non-Chinese characters.

Chinese users are very active in using emoticons. To meet the requirements of users, many IMEs (Input Method Editor) and social networking tools have created their own emoticon set. We integrate the emoticons from Sina and several popular IMEs like Sogo and build our emoticon dictionary, which contains 2919 emoticons. The punctuations are extracted by the same way. There are 227 punctuations in total. The acronym list consists of English characters and digits. It includes English words such as “good”, “you”, “love”, internet slangs such as “hoho”, “MM”, and abbreviations such as “OMG”, “CEO”. The feature set size for acronym is 29289.

The usage of stylistic features are different from those in previous studies in that we use them as tokens as word unigrams rather than the total number of occurrences.

4.2 Lexical Features

We represent each user’s tweets by a vector of word frequencies. We use the ICTCLAS tool¹ to segment each tweet into words. The vocabulary size for word unigram in our experiment is 158910. We do not remove stop word as in text categorization. This is because some of the stop words are actually discriminative.

4.3 Syntactic Features

Since the ICTCLAS software also outputs the POS tags for each word, we then use these tags as syntactic features. Figure 2 shows the syntactic structures of two sample sentences. One is correct, and the other is wrongly segmented.



Fig. 2. Syntactic structure of two sample sentences

After segmentation, each word has a POS tag. For example, “father” is tagged as a “n” standing for “noun”, and “always” to be a “d” for “adverb”. The total number of syntactic features is 94.

¹ <http://ictclas.nlp.ir.org/>

4.4 Feature Selection and Combination

We use the χ^2 statistic as a metric to select features [21]. The χ^2 statistic can be used to measure the lack of independence of a token t and a class label c . It is defined as: $\chi(t, c) = \frac{N*(AD-CB)^2}{(A+B)(A+C)(B+D)(C+D)}$, where N is the total number of documents (users in our case), A is the number of times t and c co-occur, B is the number of times t occurs without c , C is the number of times c occurs without t , and D is the the number of times neither t nor c occurs.

We also do feature combinations by 1) merging different types of features into one long vector, and 2) ensembling the classifiers built from different types of features.

5 Experimental Evaluation

In this section, we evaluate the proposed approach. We first introduce the experiment setup, and then present the results using different settings.

5.1 Experiment Setup

All our experiments use the *SVM^{multiclass}* classifier² with default parameter settings. The data are randomly split into two parts: 80% for training and the rest 20% for test. The results are averaged over the 5-fold cross validations. We report the classification accuracy as the evaluation metric.

5.2 Experimental Results

Below we will show our experimental results.

Effects of Feature Selection

In order to compare the effects of feature selection, we compute the χ^2 value for all features in each class of training data, and then sort the features in the descending order of their χ^2 statistic. For training and test data, we only keep features ranked high of the χ^2 list and discard those in the tail. In Figure 3, we show the effects of feature selection with a decreasing ratio of features kept. Note that a 100% setting means using all the original features, i.e., no feature selection is done.

We have the following observations from Figure 3.

- Feature selection has very positive effects on age detection for Chinese users. It can greatly improve the accuracy for all kinds of features. The accuracy increases very fast when removing the most ambiguous features and then goes steady. This finding is new and contradicts to those in existing study in Western language [15]. In the following, we will use 50% as our default setting.

² http://www.cs.cornell.edu/People/tj/svm.light/svm_multiclass.html

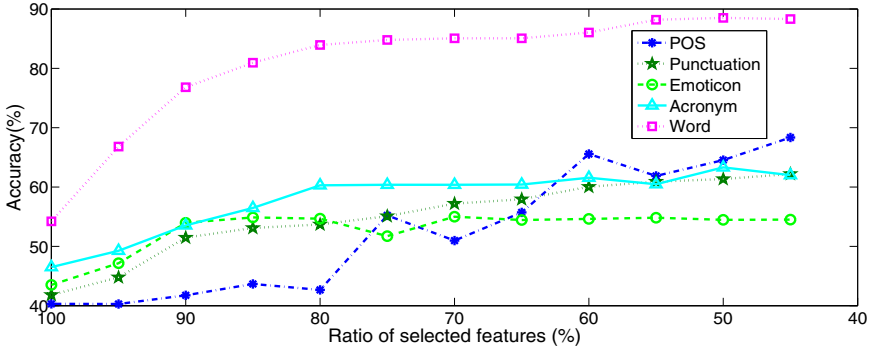


Fig. 3. The effects of feature selection

- The lexical features perform the best among all features, whether or not applying the feature selection procedure. The performances for punctuation and emoticon are generally in the middle. The accuracy for POS fluctuates when the ratio changes. At the start point, its performance is the worst. However, it increases very fast and finally outperforms all others kinds of features except the word unigram.

Effects of Feature Combinations

The effects of feature combinations are shown in Table 1. For easy comparison, we also show the accuracy for single feature on the right side.

Table 1. Effects of feature combination

Combined Features	Acc(longvec)	Acc(ensemble)	Single Feature	Acc
Emotion_Punt_Acronym	65.49	65.56	Acronym	63.31
Emotion_Punt_Acronym_Word	86.81	85.19	Punt	61.34
Emotion_Punt_Acronym_POS	68.66	68.31	Emoticon	54.47
Word_POS	88.72	88.47	POS	64.53
ALL	87.12	85.44	Word	88.50

From Table 1, we can see that:

- Merging features into a long vector is a bit better than ensembling the outputs of classifiers from different features. In the following, we will use feature merge as the default setting for feature combination.
- While other features benefit from the merge or ensemble, the accuracy for word unigram is lower down when it is combined with stylistic features. Even if the combination of word and POS improves the performance a little, the accuracy for all features is lower down when stylistic features join. This indicates that the complicated stylistic structure is not good itself for the combination process.

Effects of the Size of Training Data

We evaluate the effects of number of training data. Figure 4 shows the results.

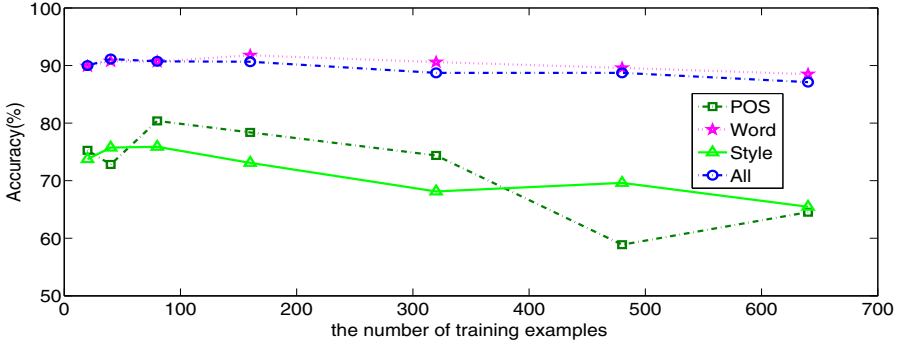


Fig. 4. The effects of training set size

In Fig. 4, we see that the accuracy for word and all features are less sensitive to the number of training examples as the stylistic (style) and syntactic (POS) features do, showing that the training set size has fewer impacts on the features whose dictionary is large. We also notice that most of the classifiers get the best result when using about 160 training examples. This can be due to the fact that the labeled set contains more noises when more users are added.

One Age Group vs. the Other Group

In Table 2, we shows the results for binary classification, i.e., using two age spans as two classes.

Table 2. Results for binary classification

	C1 vs. C2	C1 vs. C3	C1 vs. C4	C2 vs. C3	C2 vs. C4	C3 vs. C4
POS	50.13	65.50	61.44	64.63	67.25	49.94
Punt	64.00	74.44	72.25	65.94	72.06	50.69
Emoticon	54.63	56.38	68.00	54.38	65.00	60.81
Acronym	72.44	80.06	82.50	62.38	70.81	56.38
Word	64.94	81.19	88.13	69.69	75.06	50.56

In general, C1(18..22) VS. C4(48..52) get the best result. This is easy to understand because they have the biggest age difference. However, the highest accuracy for POS is achieved on C2(28..32) VS. C4(48..52). This is consistent with our observation that the young adults show specific syntactic patterns, as we will show in Fig. 7. In addition, the accuracy for C3(18..22) VS. C4(28..32) is the worst, showing that the adults and old people share more common language patterns than others do.

6 Age Related Language Patterns

In this section we analyze some interesting language patterns.

6.1 Lexical Patterns

We find significant topic divergence among the four age groups in spite of the segmentation errors. Figure 5 shows the top 10 discriminative word unigrams for each group. From the table, we have the following important notes:

1. The young people seem to be polarized. On one hand, they are interested in those close to their life such as “service” or “experience”. On the other hand, the political affairs are still their major topics. For example, “China” and “people” appear as two of the top 10 words in this class.
2. The young adults focus mainly on their daily life. The top three words are about their work status, and none of the top 10 words is related with public affairs. The reason may be that the folks in this age just start to work and many of them are going to get or already got married. They do not have enough spare time.
3. The adult people are the main stream in society. Their interests totally devote to their career. As can be seen, all the top 10 patterns are on the business or marketing.
4. The old people pay more attentions on authorities and abstractions. Eight of ten words are related to the identity which reflects one’s social status such as “civilian”, “cadre”, “chairman”, and “scholar”.

Young ^o	Young Adult ^o	Adult ^o	Old ^o
公司 <i>company</i> ^o	加班 <i>work overtime</i> ^o	行业 <i>industry</i> ^o	百姓 <i>civilian</i> ^o
服务 <i>service</i> ^o	上班 <i>go to work</i> ^o	营销 <i>advertise</i> ^o	干部 <i>cadre</i> ^o
体验 <i>experience</i> ^o	下班 <i>off work</i> ^o	商业 <i>commerce</i> ^o	书记 <i>party secretary</i> ^o
中国 <i>China</i> ^o	哎 <i>sigh</i> ^o	消费 <i>consumption</i> ^o	中央 <i>government</i> ^o
工作 <i>work</i> ^o	果然 <i>as expected</i> ^o	案例 <i>case</i> ^o	主席 <i>chairman</i> ^o
进行 <i>ongoing</i> ^o	貌似 <i>look like</i> ^o	推广 <i>promotion</i> ^o	学者 <i>scholar</i> ^o
北京 <i>Beijing</i> ^o	伤不起 <i>like a dream</i> ^o	互联网 <i>internet</i> ^o	现代 <i>modern</i> ^o
人民 <i>people</i> ^o	哈哈 <i>haha</i> ^o	业务 <i>business</i> ^o	发表 <i>publish</i> ^o
及 <i>and</i> ^o	肿么 <i>what</i> ^o	运营 <i>operation</i> ^o	教授 <i>professor</i> ^o
功能 <i>function</i> ^o	婚礼 <i>wedding</i> ^o	智慧 <i>wisdom</i> ^o	主任 <i>director</i> ^o

Fig. 5. Top 10 discriminative word unigrams

6.2 Stylistic Patterns

The most discriminative acronyms for age groups are shown in Table 3. Besides the topic divergence as in lexical patterns, we find that the young people are active in using foreign or new Internet slangs. For example, “get” usually means one masters a new skill and “QAQ” stands for a sad expression. Please also note

Table 3. Discriminative acronyms for age groups

Young	Young Adult	Adult	Old
get	ing	CEO	bull
po	mark	App	it
LOL	TM	GDP	like
QAQ	ps	pptx	is
come	pose	HR	on

that “LOL” is a game name in Chinese in most of the cases instead of “laugh out loudly” in English. Meanwhile, we find that other three groups all use acronym in a regular way.

The emoticon patterns are shown in Figure 6. We do not observe obvious patterns for old people. This indicates that the old ones are conservative and tend to use conventional emoticons. In contrast, the young people are willing to accept new and vivid emoticons such as “doge” and “shy”. Furthermore, while young people are fond of use emoticons for self-expression, young adults are more concerned with the others. For example, they use “V5” and “handsome” to compliment or praise other people.












Young ^o	Young Adult ^o	Adult ^o	Old ^o
 [doge] ^o	 [hehe] ^o	 [cake] ^o	- ^o
 [oh yeah] ^o	 [V5] ^o	- ^o	- ^o
 [love] ^o	 [money] ^o	- ^o	- ^o
 [laugh to tears] ^o	 [snow] ^o	- ^o	- ^o
 [shy] ^o	 [handsome] ^o	- ^o	- ^o

Fig. 6. Emoticon patterns

Token ^o	Syntax ^o	Examples ^o
<i>e</i> ^o	叹词 <i>exclamation</i> ^o	哦、嗨、啊 ^o
<i>al</i> ^o	形容词性惯用语 <i>adjective phrase</i> ^o	不得了，出神入化，神清气爽 ^o
<i>tg</i> ^o	时间词性语素 <i>time related phrase</i> ^o	为止，及，来着 ^o
<i>y</i> ^o	语气词 <i>interjection</i> ^o	么、呢、吧 ^o
<i>qv</i> ^o	动量词 <i>special purpose verb</i> ^o	次、回、趟 ^o

Fig. 7. Special syntactic patterns for young adults

6.3 Syntactic Patterns

While the other three groups are similar in using of syntactic structures, we find the young adults show special POS patterns. The most distinctive patterns for

this group are shown in Figure 7. We notice the top two patterns are exclamation and adjective phrases, and the others are about time and unit. We believe this corresponds to the special topic for this group. Remember that they are overwhelmingly related to daily life.

7 Conclusion

In this paper, we investigate the age detection problem for Chinese users. We construct an annotated corpus using rule based filtering and manual check. We extract three types of features to represent users, namely, stylistic, lexical, and syntactic features. We find the word unigrams are the most discriminative features in detecting age, and the syntactic and stylistic features are also informative. The improvement of feature selection is evaluated as significant on all types of features, which contradicts to existing study in Western languages. Our research also discloses a number of interesting language patterns specific to a particular age group. The results do provide us important insights in terms of analyzing the relationship between the Chinese language and their users at various ages.

Our current study focuses on classifying Chinese users into social age group. In the future, we plan to extend our work by detecting age as a continuous variable and predicting users' life stages. In addition, as a first attempt, we only explore three basic types of features. The impacts of other features need further investigation. Finally, our current experiment only involves users' own information. Our next work will study how the users in social network are affected by their neighbors.

Acknowledgments. The work described in this paper has been supported in part by the NSFC projects (61272275, 61232002, 61202036, 61272110, and U1135005), the 111 project(B07037), and SRFDP (20120141120013).

References

1. Bergsma, S., Durme, B.V.: Using conceptual class attributes to characterize social media users. In: Proc. of ACL, pp. 710–720 (2013)
2. Cheng, N., Chen, X., Chandramouli, R., Subbalakshmi, K.P.: Gender identification from e-mails. In: CIDM, pp. 154–158 (2009)
3. Garera, N., Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proc. of ACL and IJCNLP, pp. 710–718 (2009)
4. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Proc. of ICWSM, pp. 214–217 (2009)
5. Gressel, G., Hrudya, P., Surendran, K., Thara, S., Aravind, A., Poornachandran, P.: Ensemble learning approach for author profiling. In: PAN at CLEF (2014)
6. Kabbur, S., Han, E.H., Karypis, G.: Content-based methods for predicting web-site demographic attributes. In: Proc. of ICDM (2010)
7. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. PNAS **110**, 5802–5805 (2013)

8. Li, J., Ritter, A., Hovy, E.: Weakly supervised user profile extraction from twitter. In: Proc. of ACL, pp. 165–174 (2014)
9. Mislove, A., Viswanath, B., Gummadi, P.K., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proc. of WSDM, pp. 251–260 (2010)
10. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: Proc. of EMNLP, pp. 207–217 (2010)
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: Proc. of ICWSM, pp. 439–448 (2013)
12. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proc. of the 5th ACL-HLT Workshop, pp. 115–123 (2011)
13. Nguyen, D., Trieschnigg, D., Doğruöz, A.S., Grave, R., Theune, M., Meder, T., de Jong, F.: Why gender and age prediction from tweets is hard: lessons from a crowdsourcing experiment. In: Proc. of COLING, pp. 1950–1961 (2014)
14. Otterbacher, J.: Inferring gender of movie reviewers: exploiting writing style, content and metadata. In: Proc. of CIKM, pp. 369–378 (2010)
15. Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting age and gender in online social networks. In: Proc. of SMUC, pp. 37–44 (2011)
16. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proc. of SMUC, pp. 37–44 (2010)
17. Rosenthal, S., McKeown, K.: Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In: Proc. of ACL, pp. 763–772 (2011)
18. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp. 199–205 (2005)
19. Tam, J., Martell, C.H.: Age detection in chat. In: Proc. of ICSC, pp. 33–39 (2009)
20. Xiao, C., Zhou, F., Wu, Y.: Predicting audience gender in online content-sharing social networks. *JASIST* **64**, 1284–1297 (2013)
21. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proc. of ICML, pp. 412–420 (1997)