Cross-media Age Regression with Textual Adaptation

Jing Chen¹, Long Cheng¹, Bo Deng¹, Jun Liang¹, Ji Qi¹ and Shoushan Li²,a

¹ China Mobile (Suzhou) Software Technology Co., LTD, 215000 Suzhou Jiangsu, China
² College of Computer Science, Soochow University, 215000 Suzhou Jiangsu, China

a Corresponding author: lishoushan@suda.edu.cn

Abstract. In realistic scenarios, an age regression model learned from one social media (named the source media) generally performs rather poorly when it is tested on another social media (named the target media). In this paper, a textual adaptation approach is proposed to cross-media age regression which aims to improve the age regression performance by exploiting textual features in the labeled data from the source media and unlabeled data from the target media. The basic idea to achieve this lies in the fact that many textual features are shared by the data from both social media. Specifically, two views generated by random subspace generation (RSG) are leveraged to train two separate regressors in a co-training algorithm for adding automatically-labeled samples in the target media. Moreover, we tackle the confidence evaluation challenge in co-training by the query by committee (QBC) approach. Empirical studies demonstrate the effectiveness of the proposed approach to cross-media age regression.

1. Introduction
In social analysis, one fundamental task is known as age prediction which aims to predict the ages of online users. Recent years have witnessed an exploding interest in this task and substantial research efforts have been conducted on this research issue from several research communities, such as natural language processing [1] and social network [2]. Meanwhile, age prediction has been an essential pre-processing step in many social applications and accurate age prediction of social media users from evidence benefits these applications, such as intelligent marketing [3, 4], online advertising [5], and personality analysis [6]. However, almost all the previous studies perform age prediction in a single social media. In some real scenarios, several social media may be involved and not all of them have enough labeled data to train a well-performed predictor in each media. For example, Facebook.com is a social media site where many users publicize their ages in their homepages, making the collection of labeled data easy, while Linkedin.com is a social media site where age information is not available in users’ homepages, making the collection of labeled data difficult. This scenario raises a novel task concerning age prediction, which aims to leverage labeled data from a social media to train a prediction model for testing data from a different social media. Briefly, we refer to this novel task as cross-media age prediction where the social media with labeled data are called the source media and the other media are called the target media.

| Name: **** | Age: 23 |
|----------------|-----------|
| Social Information (e.g., User information) |
Followings:
- Zhuoer Xu (ID: 2139500310),
  Uncle Pan (ID: 2101065794),
  Cjjj77 (ID: 305165914)......

Textual Information (e.g., Messages)

(1) UGH I don’t wanna go to school tomorrow.
    DONT WANNA SEE A TEACHERFACE
    AGAIN, OH WAIT I’M 23! WAHOOOOO!!

(2) WHAT TIME IS IT?! GAME TIME!

Figure 1. A user example in SINA social media

Name: ****  Age: 23

Social Information (e.g., User information)

Followings:
- Xiao Yi (ID: 55d0aa2ce4bo003f1952c97),
  Huihui Gu (ID: 55d0a3ee4b003f195303f),
  Cjjj77 (ID: 2cfc636a6a6a3737c039)......

Textual Information (e.g., Messages)

(1) I want to show out that I have no homework today, HAHAHAHA.
(2) Then I am thinking that should I update the address from my home to my school.

Figure 2. A user example in TIEBA social media

Cross-media age regression is challenging and many existing approaches to age prediction are not readily feasible for this novel task. For instance, social information is essential in some age prediction approaches, such as the approaches by Rosenthal [1] and Goswami et al. [2]. Even worse, many social features vary in different social media. Some social features are no longer helpful when they are utilized in a different social media site. For example, Figure 1 shows an online user from SINA social media and Figure 2 shows an online user from TIEBA social media. From the two figures, we can see that the IDs of following users in each site are totally different, even though they refer to the same person, e.g., “cjjj77 (ID: 305165914)” in SINA and “cjjj77 (ID: 2cfc636a6a6a3737c039)” in TIEBA. Therefore, in cross-media age regression, the features of following IDs become useless in those modes that employ following IDs as crucial features.

In this paper, we propose a textual adaptation approach to cross-media age prediction. The basic idea behind our approach is based on the fact that the user-generated text shares a lot of features for inferring user ages, even though the users are from different social media. For example, in Figure 1, the user publishes a message “UGH I don’t wanna go to school tomorrow,” which contains the features such as “school” for inferring the user to be a student’s age. In Figure 2, the user publishes a message “Then I am thinking that should I update the address from my home to my school” which contains some similar key features. If we know the user in SINA is a student, it is easy to infer the user in TIEBA to be a student, so as to limit his/her age in a suitable range.

However, the performance of the textual adaptation approach may remain limited due to the fact that the textual features in different social media sites are sometimes differently distributed. To alleviate this problem, we present a semi-supervised approach to exploit the unlabeled data from the

---

1 http://weibo.com/
2 http://tieba.baidu.com/
target media to balance the different feature distributions in the two media. Our semi-supervised approach is mainly based on the famous semi-supervised learning approach namely co-training [7], wherein the Random Subspace Generation (RSG) approach [8] is applied to generate multiple feature views.

It is also worth to note that one challenge in this approach is that the prediction confidence of a unlabeled sample used in the co-training algorithm is not easy in our regression problem. Co-training is originally applied in the classification problem where estimating the prediction confidence on a unlabeled sample could naturally use the classification posterior probabilities. However, in regression, no such posterior probabilities are provided. To tackle this challenge, we first applied RSG approach again by generating several random feature subspaces and then use the query by committee (QBC) method, together with the generated multiple feature subspaces, to measure the labeling confidence of each unlabeled instance.

2. Related Work
This section gives an overview of related work on age classification and age regression respectively.

2.1 Age Classification
Over the last decade, the majority of studies model age prediction as a classification problem and conduct their empirical studies on two main domains, i.e., blog [9, 10] and social media [11].

In the blog domain, Schler et al. [10] focus on textual features extracted from the blog text, such as word context features and parts-of-speech (POS) stylistic features. Burger and Henderson [9] explore some social features, such as location, time, and friend features, related to blogger ages. Ikeda et al. [12] is one exception which proposes a semi-supervised approach to age prediction by training multiple sub-classifiers with textual features. Some other studies, such as Rosenthal [1] and Goswami et al. [2], explore both the textual and social features in automatic age classification.

In the social media domain, Mackinnon and Warren [11] explore some kind of social features, i.e., the relationship between users to predict a user’s age and country of residence in a social network. Peersman et al. [13] apply a text categorization approach to age classification with textual features only. More recently, Marquardt et al. [14] propose a multi-label classification approach to predict both the gender and age of authors from texts.

2.2 Age Regression
Compared to age classification, much less studies model age prediction as a regression problem.

Nguyen et al. [15] explore textual features, such as word unigrams, POS unigrams and bigrams, together with gender features in age regression via a linear regression model. Their empirical studies find that word unigrams can achieve reasonable performance and that POS patterns are strong indicators of the old age. Nguyen et al. [16] further explore age prediction of Twitter users with a linear regression model. They find that an automatic system can achieve better performance than a human being. Chen et al. [17] apply active learning approach to age regression for better exploiting the unlabeled data to improve the performance. Chen et al. [18] explore deep learning approach to age regression and combine results of age regression and classification linearly to improve the performance of age prediction.

However, no previous studies focus on cross-media age regression. To the best of our knowledge, this is the first attempt to address the age regression task when two different social media are involved.

3. Background
In this section, we give some background on data collection and the Random Sub-space Generation (RSG) approach.

3.1 Data Collection
Our data is collected from SINA Weibo and TIEBA, two famous social media in China. SINA Weibo is a microblog platform, and TIEBA is a keyword-based online forum. The information of users can be acquired through analyzing their profiles and posts.

From SINA, we crawl each user’s homepage which contains user information (e.g. name, age, gender, followings, verified type) and their posted messages. The data collection process starts from some randomly selected users, and iteratively gets the data of their follower and followings. We remove those unsuitable users who are verified as organizations because the age attributes of these users make no sense. From TIEBA, we crawl each user’s homepage which contains user information (e.g. name, gender, birthday, followings, blood type) and their posted messages. Besides, although the posted messages are the basic and major factor to predict user ages [15], some users post very few messages. To guarantee the reliability of the data, we remove those not especially active users who post less than 50 messages in each social media. In total, we collect the homepages of about 12000 users in SINA and 11000 users in TIEBA.

![Figure 3](image1.png)

**Figure 3. User distribution of different ages in SINA**

![Figure 4](image2.png)

**Figure 4. User distribution of varying ages in TIEBA**

Figure 3 and figure 4 shows the user distribution in different ages in SINA and TIEBA. From the two figures, we can see that the data distribution of user ages either from SINA or from TIEBA is rather imbalanced. Most users are young whose ages are in the range of 19-28 in SINA and in the range of 19-32 in TIEBA.
Table 1. Statistics about the unique word unigram features

|       | $F_{SINA}$ | $F_{TIEBA}$ | $|F_{SINA} \cap F_{TIEBA}|$ |
|-------|------------|-------------|-----------------------------|
| Number of features | 281678 | 410973 | 62520 |

Table 1 shows the statistics about the textual features, i.e., unique word unigram features, from the two social media, where $F_{SINA}$ denotes unigram features from the SINA data, and $F_{TIEBA}$ denotes unigram features from the TIEBA data, and $|F_{SINA} \cap F_{TIEBA}|$ denotes the joint set of $F_{SINA}$ and $F_{TIEBA}$. From this table, we can see that the feature sets from the two social media share 62520 word features which make it possible to perform age regression from one social media to the other social media.

3.2 Random Feature Subspace

Random Subspace Generation (RSG) is a popular approach to generating multiple learners trained with several feature subspaces [8]. Assume $L = (x_1, x_2, \ldots, x_n)$ the training data and $x_i$ an $m$-dimensional vector $x_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, described by $m$ features. RSG first randomly selects $r$ ($r < m$) features and obtains an $r$-dimensional random subspace of the original $m$-dimensional feature space. In this way, a modified training set $L' = (x'_1, x'_2, \ldots, x'_n)$ consisting of $r$-dimensional samples $x'_i = (w'_{i1}, w'_{i2}, \ldots, w'_{ir})$ is generated. Then, a subspace regression learner can be trained in random subspaces $x'$ using the modified training set.

In age regression, we split all textual features into $N$ disjoint feature subsets which are used as feature subspaces. Thus, the dimension of the subspace $r$ is $m/N$.

4. Cross-media Age Regression

In cross-media age regression, the test samples are from the target media while the training samples are from the source media. In the sequel, the labeled data from the source media is denoted as $L_s = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^m$ is the m-dimensional input vector, and $y_i$ is its label. The unlabeled data from the target media is denoted as $U_t = \{(x'_i)\}_{i=1}^n$ where $x'_i \in \mathbb{R}^m$ is the m-dimensional input vector.

Cross-media age regression is the task of leveraging $L_s$ and $U_t$ to train a regressor $f$ for predicting samples from the target media. For clarity, some important symbols are summarized in Table 2.

Table 2. Symbol definition

| Symbol | Definition |
|--------|------------|
| $L_s$  | Labeled data from the source media |
| $L_t$  | Automatically-labeled data from the target media |
| $U_t$  | Unlabeled data from the target media |
| $f$    | The trained regressor |
| $L'$   | Subspace labeled data with the $r$-dimensional feature subspace |

4.1 Self-training for Cross-media Age Regression

Self-training is a straightforward approach to semi-supervised learning [19]. The main idea of this approach is to firstly train a classifier with a small amount of labeled data, and then iteratively retrain it by adding most confident unlabeled samples as new labeled data. Instead of training a classifier, we need to train a regressor in our task.

In this study, we utilize some labeled data from the source media, and increase the amount of labeled data using the unlabeled data from the target media to improve the performance of cross-media...
age regression with textual information by self-training algorithm. Figure 5 shows the self-training algorithm for cross-media age regression. In this algorithm, one remaining challenge is the confidence measurement method used in Step (3), which will be solved in Section 4.3.

**Input:**
- \( L_s \): Labeled data in the source media
- \( U_T \): Unlabeled data in the target media

**Output:**
- \( L_T \): Automatically-labeled data in the target media

**Procedure:**
1. Initialize \( L_T = \emptyset \)
2. Loop until \( U_T = \emptyset \)
   1. Learn regression function \( f \) with \( L_s \)
   2. Use \( f \) to label the samples from \( U_T \)
   3. For each age category:
      - Choose one sample \( t_i \) which is most confidently predicted by \( f \)
   4. \( L_T = L_T + t_i \), \( U_T = U_T - t_i \)

Figure 5. Self-training algorithm for cross-media age prediction

### 4.2 Co-training for Cross-media Age Regression

The co-training algorithm is a famous semi-supervised learning approach which starts with a small set of labeled data and increases the amount of labeled data using the unlabeled data by bootstrapping with two views [7].

In this study, we utilize some labeled data from the source media and increase the amount of labeled data using the unlabeled data from the target media to improve the performance of cross-media age regression with textual information by co-training algorithm.

The two views in our co-training algorithm are generated from two feature subspaces. Specifically, we use RSG to generate two feature subspaces by randomly splitting the whole textual features into two disjoint subspaces (i.e., \( N = 2 \)). Thus, the labeled data from the source media, i.e., \( L_s \) generates two subspace data sets, which are denoted as \( L_1^s \) and \( L_2^s \). Meanwhile, the unlabeled data from the target media, i.e., \( U_T \) generates two subspace data sets, which are denoted as \( U_1^T \) and \( U_2^T \).

**Input:**
- \( L_1^s \): Labeled data of the first view in the source media
- \( L_2^s \): Labeled data of the second view in the source media
- \( U_1^T \): Unlabeled data of the first view in the target media
- \( U_2^T \): Unlabeled data of the second view in the target media

**Output:**
- \( L_1^T \): Automatically-labeled data of the first view in the target media
- \( L_2^T \): Automatically-labeled data of the second view in the target media

**Procedure:**
1. Initialize \( L_1^T = \emptyset \) and \( L_2^T = \emptyset \)
2. Loop until \( U_1^T = \emptyset \) or \( U_2^T = \emptyset \)
1) Learn regression function \( f^1 \) with \( L^1_L + L^2_L \)
2) Learn regression function \( f^2 \) with \( L^1_L + L^2_L \)
3) Use \( f^1 \) to label the samples from \( U^1_T \)
4) Use \( f^2 \) to label the samples from \( U^2_T \)
5) For each age category:
   5.1) Choose one sample \( t^1_r \) of the first view which is most confidently predicted by \( f^1 \)
   5.2) Choose the corresponding samples \( t^2_r \) of the second view (the sample from the same user as \( t^1_r \))
   5.3) Choose one sample \( t^2_r \) of the second view which is most confidently predicted by \( f^2 \)
   5.4) Choose the corresponding samples \( t^1_r \) of the second view (the sample from the same user as \( t^2_r \))
6) \( L^1_T = L^1_L + t^1_r + t^2_r \)
   \( L^2_T = L^2_L + t^1_r + t^2_r \)
7) \( U^1_T = U^1_T - t^1_r - t^2_r \)
   \( U^2_T = U^2_T - t^1_r - t^2_r \)

Figure 6. Co-training algorithm for cross-media age prediction

Figure 6 shows the co-training algorithm for cross-media age regression. In this algorithm, the same challenge to co-training algorithm for cross-media age prediction is the confidence measurement method used in Step (5.1) and Step (5.3), which will be solved in Section 4.3.

4.3 Confidence Measurement with QBC and RSG

Originally, query by committee (QBC) is a group of active learning approaches which employ a committee of learners to select an unlabeled sample at which their classification predictions are maximally spread [20].

Different from active learning, semi-supervised learning aims to select the most certain sample rather than the most uncertain sample. Thus, we select an unlabeled sample at which their regression predictions are maximally agreed. Formally, given the regression results from the committee of learners \( \{y^1_1, y^1_2, ..., y^q_q\} \), the confidence score is calculated as follows:

\[
\text{conf}(x) = -\log \left( \sum_{i=1}^{q} (\bar{y} - y_i)^2 \right) 
\]

(1)

Where \( \bar{y} \) is the final decision result of the committee, calculated as follows:

\[
\bar{y} = \frac{1}{q} \sum_{i=1}^{q} y_i 
\]

(2)

The more the confidence score is, the more confidently the sample is predicted.

Input:

\( L^1 : \) Labeled data of the first view \( L^1_L = L^1_L + L^2_L \)
\( U^1_T : \) Unlabeled data of the first view from the target media

Output:

The most confidently predicted sample

Procedure:

1) Adopt RSG to generate \( M \) subspace training data \( \{ L^{1_u}, L^{2_u}, ..., L^{M_u} \} \) from \( L^1 \)
2) Learn \( M \) regression functions \( \{ f^{1_u}, f^{2_u}, ..., f^{M_u} \} \) with the obtained subspace training data.
3) Use all regression functions to label the samples from $U_r$

4) Calculate and sort the confidence scores of all unlabeled samples with formula (1)

5) Pick the sample with the maximum confidence score as the most confidently predicted sample in $U_r$

Figure 7. The algorithm of confident sample selecting by QBC with RSG

To generate a committee of learners, we once again adopt the RSG approach to generate several feature subspaces for training multiple learners.

Figure 7 shows the algorithm of our approach based on QBC and RSG for selecting confident samples. Note that we only give the algorithm description on the first view. A similar description is obvious for the second view. For self-training, the optimization idea is identical except that we need not to split textual features into two disjoint subspaces as two views. Confident sample selecting is explored on whole textual features.

5. Experimentation

In this section, we have systematically evaluated our approach to cross-media age regression.

5.1 Experimental Settings

Data Setting
The data collection has been introduced in Section 3.1. To enforce task difficulty, we focus on the age range from 19 to 28, totally 10 age categories and extract 2000 samples from each social media, i.e., SINA and TIEBA. This data contains a balanced data set in each age, i.e., 200 samples in each age.

In single-media age regression, we use 80% of the data in each age category as the training data and 20% of the data as the test data. In the training data, 10 users are randomly selected as initial labeled data and the remaining users are used as unlabeled data. This setting is designed to test our approach as a semi-supervised learning approach in a single media wherein the training and test data are from the same media.

In cross-media age regression, we use 80% data from the source media in each age category as the labeled data, and 80% data from the target media in each age category as the unlabeled data and the other 20% data in the target media as the test data.

Regression Algorithm & Features
We use the libSVM$^3$ tool to implement our SVM regression algorithm with the linear kernel and the textual features. The parameters are set to the default value.

Evaluation Metric
We employ the coefficient of determination $R^2$ to measure the regression performance. Coefficient of determination $R^2$ is used in the context of statistical models with the main purpose to predict the future outcomes on the basis of other related information. $R^2$ is a number between 0 and 1. $R^2$ nearing 1.0 indicates that a regression line fits the data well [21].

5.2 Experimental Results
For thorough comparison, we implement following several approaches to cross-media age regression:

Baseline: which directly trains the labeled data from the source media to test the samples from the target media (no unlabeled data in the target media is used.)

Self-training+KNN: which employs self-training to pick automatically-labeled samples in the unlabeled data from the target media. The self-training is the same as our approach which is illustrated in Section 4.1. However, the confidence is measured by the KNN approach [22]. The KNN approach first calculates each sample’s K nearest neighbors in the labeled data and then regards the sample with the largest number of the neighbors in the same age as the most confident sample. The estimated label of the picked sample is also decided by the KNN approach. Here, K is set to be 10.

---

$^3$ http://www.csie.ntu.edu.tw/~cjlin/libsvm/
Co-training+KNN: which employs co-training approach pick automatically-labeled samples in the unlabeled data from the target media. The co-training is the same as our approach which is illustrated in Section 4.2. However, the confidence measurement method is KNN approach as discussed aforementioned.

Self-training+QBC: which employs self-training to pick automatically-labeled samples from the unlabeled data in the target media, as illustrated in Section 4.1. The confidence is measured by our QBC approach as discussed in Section 4.3. Here, the subspace number M is set to be 4.

Co-training+QBC: which employs co-training approach pick automatically-labeled samples from the unlabeled data in the target media, as illustrated in Section 4.2. The confidence is measured by our QBC approach as discussed in Section 4.3. Here, the subspace number M is set to be 4.

Figure 8. Performance comparison of different approaches to single-media age regression

5.2.1 Experimental Results of Single-media Age Regression
Figure 8 shows the performance of different approaches to semi-supervised age regression where the training and test data are both from the same media. From this figure, we can see that, Self-training+KNN and Co-training+KNN perform even worse than the Baseline approach, which indicates that the KNN approach is not a good choice for measuring the regression confidence in our task. Although Self-training+QBC outperforms the Baseline approach when it is tested in SINA, it fails to improve the performance when it is tested in TIEBA. Among all approaches, Co-training+QBC performs best and it performs consistently better than the other approaches either in SINA or in TIEBA. Specifically, Co-training+QBC achieves an improvement of 3.4% in SINA and an improvement of 3.2% in TIEBA compared to the Baseline approach.
5.2.2 Experimental Results of Cross-media Age Regression

Figure 9 shows the performance of different approaches to cross-media age regression. From this figure, we can see that Self-training+KNN and Co-training+KNN perform even worse than the Baseline approach, which indicates that the KNN approach for measuring the regression confidence is not a good choice in our task. Although Self-training+QBC outperforms the Baseline approach when it is tested from SINA to TIEBA, it fails to improve the performance when it is tested from TIEBA to SINA. Among all approaches, Co-training+QBC performs best and it performs consistently better than other approaches either from SINA to TIEBA or from TIEBA to SINA. Experimental results of cross-media age regression are consistent with the experimental results of single-media age regression.
Figure 10 shows the detailed performances of our Co-training+QBC approach to cross-media age regression when different sizes of unlabeled are added in the co-training process. From this figure, we can see that our approach is successful in exploiting unlabeled data in the target media and the performance is consistently improved when more unlabeled samples are added. When all unlabeled samples are used, the performances reach the best, resulting in an improvement of 3.1% from SINA to TIEBA, and an improvement of 4.8% from TIEBA to SINA compared to the Baseline approach.

In our implementation, the number of the feature subspace in each view, i.e., $M$, is an important parameter in our approach. To evaluate the sensitiveness of this parameter, we set it to 2, 4, 6, 8, 10 respectively. Figure 11 shows the performance of different feature subspaces. From Figure 11, we can see that our approach consistently performs well when $M$ is set in the range of (2, 8).
6. Conclusion
In this paper, a novel task in age regression named cross-media age regression is proposed which aims to exploit the labeled data from the source media and unlabeled data from the target media to train a regressor for predicting the samples in the target media. Our approach leverages two views generated by RSG to train two separate regressors in a co-training algorithm for automatically labeling the unlabeled data from the target media. Moreover, the confidence measurement problem in our regression model is solved by using a committee of feature-subspace regressors in each view. Evaluation shows that our approach effectively improves the performance in cross-media age regression by exploiting the unlabeled data from the target media.

In our future work, we would like to improve the performance on cross-media age regression by exploring more features and unlabeled data. Moreover, we would like to apply our approach to semi-supervised learning on cross-media regression in some other NLP tasks.

References
[1] S. Rosenthal. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre-and Post-Social Media Generations. ACL, 763-772 (2011).
[2] S. Goswami, S. Sarkar, M. Rustagi. Stylometric Analysis of Bloggers’ Age and Gender. AAAI, 214-217 (2009).
[3] D. Preotiuc-Pietro, V. Lamos, and N. Aletras. An analysis of the user occupational class through twitter content. ACL, 1754–1764 (2015).
[4] J.J. Wang, Y.Q. Zhu, D.Y. Li, et al. Joint User Attributes and Item Category in Factor Models for Rating Prediction. DASFAA, 277-296 (2016).
[5] S. Volkova, T. Wilson, D. Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. EMNLP, 1815–1827 (2013).
[6] B. O’Connor, R. Balasubramanyan, B. Routledge, et al. From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 1842-1850 (2010).
[7] A. Blum, T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. ACL, 92-100 (2000).
[8] T. Ho. The Random Subspace Method for Constructing Decision Forests. PAMI, 20(8): 832-844 (1998).
[9] J.D. Burger, J.C. Henderson. An Exploration of Observable Features Related to Blogger Age. AAAI, 15-20 (2006).
[10] J. Schler, M. Koppel, S. Argamon, et al. Effects of Age and Gender on Blogging. AAAI, 199-205 (2006).
[11] I. Mackinnon, R. Warren. Age and Geographic Inferences of the LiveJournal Social Network. ICML, 176-178 (2006).
[12] D. Ikeda, H. Takamura, M. Okumura. Semi-Supervised Learning for Blog Classification. AAAI, 1156-1164 (2008).
[13] C. Peersman, W. Daelemans, L.V. Vaerenbergh. Predicting Age and Gender in Online Social Networks. SMUC, 37-44 (2011).
[14] J. Marquardt, G. Farnadi, G. Vasudevan, et al. Age and gender identification in social media. CLEF, 1129-1136 (2014).
[15] D. Nguyen, N.A. Smith, C.P. Rose. Author Age Prediction from Text using Linear Regression. ACL, 115-123 (2011).
[16] D. Nguyen, R. Gravel, D. Trieschnigg, et al. “How Old Do You Think I Am?”: A Study of Language and Age in Twitter. AAAI, 439-448 (2013).
[17] J. Chen, S.S. Li, B. Dai, et al. Active Learning for Age Regression in Social Media. Springer International Publishing, 351-362(2016).
[18] J. Chen, S.S. Li, J.J. Wang, et al. User age prediction by combining classification and regression. Sci Sin Inform, 47: 1095–1108(2017).
[19] S. Abney. Bootstrapping. ACL, 360–367 (2002).
[20] Y. Freund, H. Seung, E. Shamir, et al. Selecting Sampling Using the Query by Committee Algorithm. Machine Learning 28, 133-168 (1997).

[21] A. Cameron, F. Windmeijer. R-squared Measures for Count Data Regression Models with Applications to Health-care Utilization. Journal of Business and Economic Statistics, 14(2): 209-220 (1993).

[22] Z.H. Zhou, M. Li. Semi–Supervised Regression with Co-Training. IJCAI, 908-913 (2005).