Cross-media User Profiling with Joint Textual and Social User Embedding

Jingjing Wang[1,2], **Shoushan Li**[1,2], **Mingqi Jiang**[1,2], **Hanqian Wu**[3], **Guodong Zhou**[1,2,*]
[1]NLP Lab, School of Computer Science and Technology, Soochow University, China
[2]Collaborative Innovation Center of Novel Software Technology and Industrialization
[3]School of Computer Science and Engineering, Southeast University, China
djingwang@gmail.com, {lishoushan, gdzhou}@suda.edu.cn
20165227010@stu.suda.edu.cn, hanqian@seu.edu.cn

Abstract

In realistic scenarios, a user profiling model (e.g., gender classification or age regression) learned from one social media might perform rather poorly when tested on another social media due to the different data distributions in the two media. In this paper, we address cross-media user profiling by bridging the knowledge between the *source* and *target* media with a uniform user embedding learning approach. In our approach, we first construct a cross-media user-word network to capture the relationship among users through the textual information and a modified cross-media user-user network to capture the relationship among users through the social information. Then, we learn user embedding by jointly learning the heterogeneous network composed of above two networks. Finally, we train a classification (or regression) model with the obtained user embeddings as input to perform user profiling. Empirical studies demonstrate the effectiveness of the proposed approach to two cross-media user profiling tasks, i.e., cross-media gender classification and cross-media age regression.

1 Introduction

User profiling is a task which leverages user generated content (UGC) to automatically identify the data about a user, such as gender (Wang et al., 2017; Li et al., 2016; Li et al., 2015) and age (Marquardt et al., 2014) identification. Recently, along with the boom of social media, user profiling has been drawing more and more attention in various social applications, such as personality analysis (Li et al., 2016; Sarawgi et al., 2011; O'Connor et al., 2010), intelligent marking (Preotiuc-Pietro et al., 2015) and online advertising (Zhang et al., 2016; Volkova et al., 2013).

In the literature, conventional approaches normally recast user profiling as a supervised learning problem (Marquardt et al., 2014; Zhang et al., 2016; Ciot et al., 2013; Corney et al., 2002) by exploring various user generated textual features and user social connection features. However, the performance largely depends on a large amount of labeled data and the performance normally degrades dramatically when the model is tested on a different social media. Even worse, many real scenarios may involve several social media with some of them lacking sufficient labeled data to train a well-performed model in each social media. For example, Facebook.com is a social media site where many users publicize their age attribute in their homepages, making the collection of labeled data easy, while Linkedin.com is a social media site where age attribute information is not available in users homepages, making the collection of labeled data rather difficult. These scenarios raise a challenging user classification task, which leverages labeled data from a social media to train a model and evaluate it on the test data from a different social media. Briefly, we refer to this task as cross-media user profiling where the social media with labeled data is called the *source* media and the social media with only unlabeled data the *target* media.

In this paper, we address cross-media user profiling by bridging the knowledge between the *source* and *target* media with a uniform user embedding learning approach. As a user representation way, user

[*] Guodong Zhou is Corresponding Author.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1410–1420
Santa Fe, New Mexico, USA, August 20-26, 2018.

User **A** from *SINA*

Gender: *female* **Age:** *22*

Social Link Information (e.g., ***followers*****)**
▷ **ID1:** *513950031*
▷ **ID2:** *205165814*

Textual Information (e.g., ***messages*****)**
(1) *Goodbye, beautiful school and dear teacher. Its time to celebrate Christmas holiday, hahahaha.*
(2) *Big surprise! So amazing Louis Vuitton's necklace for Christmas gift. Love you forever, dear boyfriend.*

User **B** from *TIEBA*

Gender: *female* **Age:** *22*

Social Link Information (e.g., ***followers*****)**
▷ **ID3:** *32d0aa2ce4*
▷ **ID4:** *7dfc636a6a*

Textual Information (e.g., ***messages*****)**
(1) *UGH i don't wanna go to school tomorrow. Don't wanna see a teacher again. Oh wait, i have been 22! Wahooooo, forgive me, God..*
(2) *What a perfect necklace! Matching my earrings so perfect!! Don't cut my hands, please, haha.*

Figure 1: The user examples from two different social media, i.e., *SINA* and *TIEBA* social media

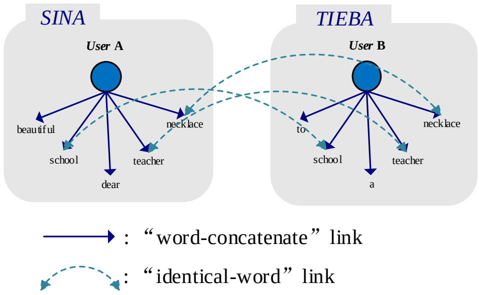


Figure 2: Cross-media user-word network

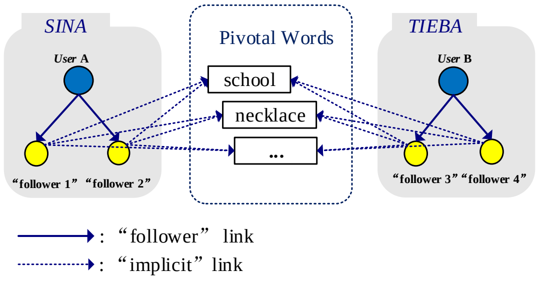


Figure 3: Cross-media user-user network with pivotal word vertices

embedding maps a user to a vector of real numbers with the motivation that embedding learning is good at capturing the relationship among the instances in the unlabeled data from both the *source* and *target* media.

One straightforward way to learn user embedding is to construct two user-word networks. Due to the phenomenon of sharing some identical words by two users, these two user-word networks are naturally connected and thus can be merged to be a mixture network, namely cross-media user-word network. For instance, Figure 1 shows two users, i.e., *User* **A** and *User* **B**, from two social media, i.e., *SINA* and *TIEBA*. It is easy to construct a user-word network wherein the users from different social media are naturally connected due to their sharing some identical words, as shown in Figure 2. Given this network, a network embedding approach, e.g., the LINE approach by Tang et al. (2015b), could be applied to learn vertex embedding and thus as user embedding.

Obviously, one main drawback of the above approach is that it much depends on the textual information and ignores the social link information which is important for user classification. Thus, a better approach to learn user embedding is to incorporate both the textual and social link information.

However, incorporating the social link information is challenging because user IDs in the different social media are totally different. As a result, the two user-user networks in the two social media are separated to each other, making it impossible to learn the relationship between two users from different social media. To tackle this challenge, we link the two user-user networks in the *source* and *target* media with some pivotal word vertices, as shown in Figure 3.

Besides, we learn better user embedding by learning a heterogeneous network composed of the two above networks, i.e., the cross-media user-word network and the cross-media user-user network. Once obtaining user embedding, we train a classifier (or regressor) on the labeled data from the *source* media

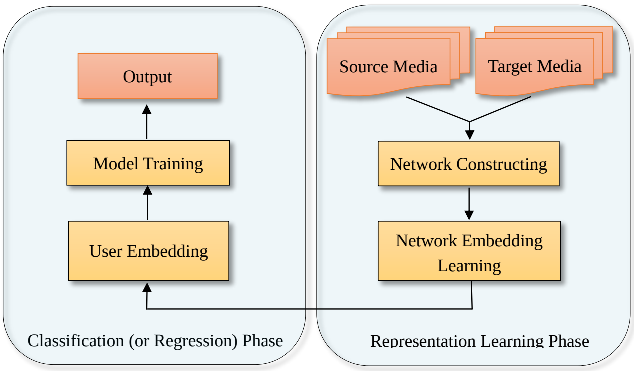

Output
Model Training
User Embedding
Classification (or Regression) Phase
Source Media
Target Media
Network Constructing
Network Embedding Learning
Representation Learning Phase

Figure 4: The overall architecture of our approach

and evaluate on the data from the *target* media. Empirical studies demonstrate that our approach outperforms both the straightforward baseline and conventional domain adaptation approaches on two different cross-media user profiling tasks, i.e., cross-media gender classification and age regression.

2 Related Work

In the last decade, many researchers have devoted their efforts on user profiling (e.g., gender identification and age identification) in several research communities, such as natural language processing and social network analysis.

For the gender identification task, Mohammad and Yang (2013) show that there are marked differences across genders in how they use emotion words in work-place email. Ciot et al. (2013) conduct the first assessment of latent attribute inference in various languages beyond English, focusing on gender inference of Twitter users. Li et al. (2015) aim to identify the genders of two interactive users on the basis of micro-blog text. Some other studies, such as Mukherjee and Liu (2010), Peersman et al. (2011) and Gianfortoni et al. (2011) focus on exploring more effective features to improve the performance. Wang et al. (2017) propose a joint learning approach in order to leverage the relationship features among relevant user attributes.

For the age identification task, most studies are devoted to explore efficient features in blog and social media. Schler et al. (2006) focus on textual features extracted from the blog text, such as word context features and POS stylistic features. Peersman et al. (2011) apply a text categorization approach to age classification with textual features extracted from the text in social media. More recently, Marquardt et al. (2014) propose a multi-label classification approach to predict both the gender and age of authors from texts adopting some sentiment and emotion features. Differently, in this paper, we cast age identification task as a regression problem instead of a classification problem.

Different from all above studies, we focus on the cross-media user profiling task. To the best of our knowledge, there are only two related papers by (Sap et al., 2014; Pardo et al., 2016) which mentions the cross-media user profiling issue. In their paper, only textual information is used in cross-media user profiling. Unlike their study, this paper firstly employs both textual and social information in cross-media user profiling.

3 Our Approach

3.1 Framework Overview

Our approach consists of two main phases, i.e., the representation learning phase and the classification (or regression) phase, which has been illustrated in Figure 4.

In the representation phase, we first construct two different types of cross-media networks, i.e., cross-media user-word network and cross-media user-user network. Second, we construct the heterogeneous user network, which is composed of the two cross-media networks. Third, we perform user embedding learning on a heterogeneous user network.

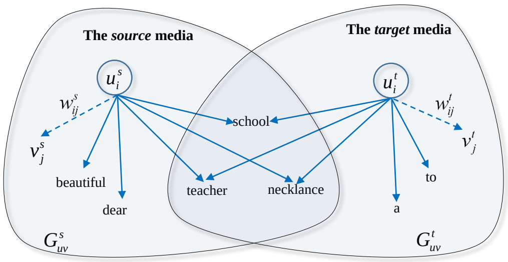


Figure 5: Cross-media user-word network G_{uv}^{cross}

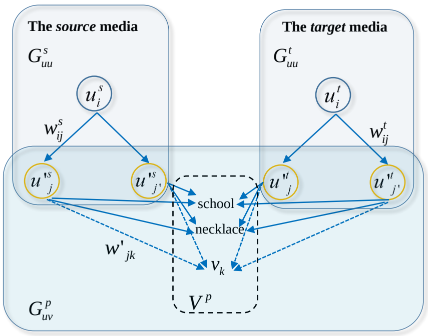


Figure 6: Cross-media user-user network G_{uu}^{cross}

In the classification (or regression) phase, we train a classifier (or regressor) with user embeddings in the *source* media for cross-media user profiling.

3.2 Basic Network Embedding Learning Approach

The goal of user embedding learning is to represent a user by a low dimensional real-value vector. In this study, we adopt the network embedding learning approach by Tang et al. (2015b) as our basic user embedding learning approach, of which we make use of the second-order proximity. Specifically, this approach is summarized as following.

Given a network $G = (V, E)$, where V is the set of vertices. E is the set of edges and the edge $e_{(i,j)}$ is directed from vertex v_i to the vertex v_j. To represent each vertex v_i with a low dimension vector $\vec{m}_i \in R^d$ where d is the dimension of this vector, we can minimize the following objective function (Tang et al., 2015b):

$$O = - \sum_{e_{(i,j)} \in E} w_{ij} \log p(v_j | v_i) \tag{1}$$

Where w_{ij} is the weight of the edge $e_{(i,j)}$. The conditional probability of vertex v_j generated by the vertex v_i, i.e., $p(v_j|v_i)$ is defined as follows:

$$p(v_j|u_i) = \frac{\exp(\vec{m}_j^\top \cdot \vec{m}_i)}{\sum_{k=1}^{|V|} \exp(\vec{m}_k^\top \cdot \vec{m}_i)} \tag{2}$$

where $\vec{m}_i$ is the embedding vector of the vertex v_i, and $\vec{m}_j$ is the embedding vector of the vertex v_j. $|V|$ is the number of all vertices in V.

3.3 Our Network Embedding Learning Approach

In our approach, we need to construct three different networks for learning network embedding, i.e., **(a)** Cross-Media User-Word Network; **(b)** Cross-Media User-User Network; **(c)** Heterogeneous User Network.

(a) Cross-Media User-Word Network Embedding

Figure 5 shows the cross-media user-word network G_{uv}^{cross}, which is composed of two user-word networks, i.e., G_{uv}^s and G_{uv}^t.

Formally, $G_{uv}^s = (U^s \bigcup V^s, E_{uv}^s)$ is a user-word network from the *source* media. It is a bipartite network where U^s and V^s are two disjoint sets of vertices, denoting the user vertices and word vertices respectively. E_{uv}^s is the set of edges between users and their published words in the *source* media.

$G_{uv}^t = (U^t \bigcup V^t, E_{uv}^t)$ is a user-word network from the *target* media. It is also a bipartite network where U^t and V^t are two disjoint sets of vertices, denoting the user vertices and word vertices respectively. E_{uv}^t is the set of edges between users and their published words in the *target* media.

For the network G_{uv}^{cross}, we can minimize the following objective function:

$$O_{uv} = -\sum_{e_{(i,j)} \in E_{uv}^s} w_{ij}^s \log p_{uv}(v_j^s|u_i^s) - \sum_{e_{(i,j)} \in E_{uv}^t} w_{ij}^t \log p_{uv}(v_j^t|u_i^t) \tag{3}$$

where the edge $e_{(i,j)} \in E_{uv}^s \; or \; E_{uv}^t$ is directed from user vertex u_i to word vertex v_j. The weight w_{ij} of the edge $e_{(i,j)}$ in both the *source* and *target* media is simply defined as the number of times v_j appears in the messages published by u_i. $p_{uv}(v_j^s|u_i^s)$ and $p_{uv}(v_j^t|u_i^t)$ are two conditional probabilities of a word vertex generated by a user vertex, which could be computed by the vertex embeddings, as shown in Equation (2).

(b) Cross-Media User-User Network Embedding

Figure 6 shows the cross-media user-user network G_{uu}^{cross}, which is a network composed of two user-user networks, i.e., G_{uu}^s and G_{uu}^t, and one user-pivotal_word network, i.e., G_{uv}^p.

Formally, $G_{uu}^s = (U^s, E_{uu}^s)$ a user-user network from the *source* media. U^s is the entire set of users in the *source* media. E_{uu}^s is the set of edges between users and their "follower" users in the *source* media.

$G_{uu}^t = (U^t, E_{uu}^t)$ is a user-user network from the *target* media. U^t is the entire set of users in *target* media. E_{uu}^t is the set of edges between users and their "follower" users in the *target* media.

As mentioned in Introduction, the user-user networks in both the *source* and *target* media share no identical users, i.e., $G_{uu}^s \bigcap G_{uu}^t = \Phi$, which makes it impossible to build connection among the users in both social media. Therefore, we attempt to construct another sub-network, a user-pivotal_word network G_{uv}^p, to bridge the users in both social media.

$G_{uv}^p = (U'^s \bigcup V^p \bigcup U'^t, E_{uu}^p)$ is a user-pivotal_word network, which is constructed to link the above two networks G_{uu}^s and G_{uu}^t. It is a bipartite network where $U'^s \subseteq U^s$ is the set of "follower" users from the *source* media and $U'^t \subseteq U^t$ is the set of "follower" users from the *target* media. E_{uu}^p is the set of edges between each "follower" user and the words in V^p which are also published by his/her following user. The basic motivation of adding these edges is based on the assumption that one user's followers tend to have similar relationship to the words that are published by this user. The detailed process of constructing the network G_{uv}^p has been illustrated in **Algorithm 1**.

For the network G_{uu}^{cross}, we can minimize the following objective function:

$$O_{uu} = -\sum_{e_{(i,j)} \in E_{uu}^s} w_{ij}^s \log p_{uu}(u_j'^s|u_i^s) - \sum_{e_{(i,j)} \in E_{uu}^t} w_{ij}^t \log p_{uu}(u_j'^t|u_i^t) - \sum_{e_{(i,k)} \in E_{uu}^p} w_{ik}' \log p_{uu}(v_k|u_i') \tag{4}$$

where the edge $e_{(i,j)} \in E_{uu}^s \; or \; E_{uu}^t$ is directed from user vertex u_i to its "follower" vertex $u_j' \in U'^s or U'^t$. The edge $e_{(i,k)} \in E_{uu}^p$ is directed from a "follower" vertex u_i' to a pivotal word vertex $v_k \in V^p$. The weights of all above edges are defined as binary values (i.e., 0 or 1). $p_{uu}(u_j'^s|u_i^s)$ and $p_{uu}(u_j'^t|u_i^t)$ are two conditional probabilities of "follower" vertex $u_j'^s$ and $u_j'^t$ generated by user vertex u_i^s and u_i^t respectively. $p_{uu}(v_k|u_i')$ is the conditional probability of pivotal word vertex v_k generated by the "follower" vertex u_i'. These three kinds of probabilities could be computed by the vertex embeddings, as shown in Equation (2).

(c) Heterogeneous User Network Embedding

The heterogeneous user network G_{joint} is composed of two cross-media networks: the cross-media user-word network, i.e., G_{uv}^{cross} and the cross-media user-user network, i.e., G_{uu}^{cross}, where the user vertices are shared across the two networks.

To learn the embeddings of the heterogeneous user network, an intuitive approach is to collectively embed the two cross-media networks, which can be achieved by minimizing the following objective function:

$$O_{joint} = O_{uv} + O_{uu} \tag{5}$$

Algorithm 1: The Network G^p_{uv} Constructing

```
Input: $U^s$: users from the source media;
       $V^s$: words from the source media;
       $U'^s$: "follower" users from the source media;
       $U^t$: users from the target media;
       $V^t$: words from the target media;
       $U'^t$: "follower" users from the target media
Output: $V^p$, $E^p_{uu}$
Initialization: $E^p_{uu} = \Phi$;
// $V^p$: words shared between source and target media;
$V^p = V^s \bigcap V^t$;
for $i = 1; i \leq |U^s|; i+ = 1$ do
    // $V_i \in V^s$: words published by $u^s_i$;
    for $v_k$ in $V_i$ do
        if $v_k \in V^p$ then
            // $U_i \in U'^s$: "follower" users of $u^s_i$;
            for $u'_j$ in $U_i$ do
                generate the edge $e_{(j,k)}$ from $u'_j$ to $v_k$;
                $E^p_{uu} = E^p_{uu} \bigcup e_{(j,k)}$;
for $i = 1; i \leq |U^t|; i+ = 1$ do
    // $V_i \in V^t$: words published by $u^t_i$;
    for $v_k$ in $V_i$ do
        if $v_k \in V^p$ then
            // $U_i \in U'^t$: "follower" users of $u^t_i$;
            for $u'_j$ in $U_i$ do
                generate the edge $e_{(j,k)}$ from $u'_j$ to $v_k$;
                $E^p_{uu} = E^p_{uu} \bigcup e_{(j,k)}$;
return $V^p$, $E^p_{uu}$;
```

We adopt the asynchronous stochastic gradient descent (ASGD) (Recht et al., 2011) using the techniques of edge sampling (Tang et al., 2015b) and negative sampling (Mikolov et al., 2013) for optimizing Equation (3) and (4). In each step, we sample a binary edge $e_{(i,j)}$ with the sampling probabilities proportional to it's edge weight w_{ij}. Meanwhile, multiple negative edges $e_{(i,j)}$ are sampled from a noise distribution $p_n(i) \propto d_i^{3/4}$ as proposed in (Mikolov et al., 2013), where d_i is the out-degree of user vertex u_i. The Equation (5) can be optimized with the textual data (the user-word network) and the social link data (the user-user network) simultaneously. We call this approach joint learning. In joint learning, both the two types of cross-media networks are used together. When the network G_{joint} is heterogeneous, the weights of the edges in the two different networks, i.e., G^{cross}_{uv} and G^{cross}_{uu} are not comparable to each other. Therefore, in our approach, we alternatively sample from the two sets of edges instead of merging all the edges together to sample as the way by Tang et al. (2015a).

4 Experimentation

4.1 Experimental Settings

Data Setting: The users are collected from *SINA*[1] and *TIEBA*[2], two famous social media in China. From these two social media, we crawl each user's homepage containing user information (e.g. *name*, *gender*,*age*), messages and "follower" users. The data collection process starts from some randomly selected users, and then iteratively gets the data of their followers. In total, we collect 10000 users from *SINA* and *TIEBA* respectively. For cross-media gender classification task, in each social media, we randomly select a balanced data set containing 3000 male users and 3000 female users. For cross-media age regression task, we focus on the age from 19-28, totally 10 age categories and extract 6000 users from each social media. This data contains a balanced data set in each age, i.e., 600 samples in each age. For

[1] https://weibo.com/

[2] https://tieba.baidu.com/index.html

both above two cross-media user profiling tasks, We randomly select 80% users in each category from the *source* media as the labeled data, 80% users in each category from the *target* media as the unlabeled data and the remaining 20% users from the *target* media as test data. Then, we consider two cases: one is to consider *TIEBA* as the *source* media and *SINA* as the *target* media, i.e., TIEBA → SINA, and the other is to consider *SINA* as the *source* media and *TIEBA* as the *target* media, i.e., SINA → TIEBA.

Representations: In our experiment, we use three different representation models to represent each user. (1) **BOW**: Each user is represented by a bag of features consisting of four types of textual features, including word unigram and two kinds of complex features, i.e., F-measure, POS-pattern, are employed. These kinds of textual features yield the state-of-the-art performance for gender classification (Li et al., 2015); To get the word, POS and parse tree features, we use the public toolkit ICTCLAS[3] to perform word segmentation, POS tagging and stanford parser[4] to perform parsing on the Chinese text. (2) **Word Embeddings**: We learn word embedding matrices of each user from training a mixed training data set consisting of labeled data in the *source* media and unlabeled data in the *target* media with skip-gram algorithm[5]. The dimensionality of word vector is set to be 200. (3) **User Embeddings**: We learn the embedding vector of each user from training three different cross-media networks, i.e., G_{uv}^{cross}, G_{uu}^{cross} and G_{joint}, which are constructed from a mixed training data set consisting of the labeled data in *source* media and unlabeled data in *target* media. For the new user vertex which not present in the training data, we get its user embedding by averaging the embedding vectors of all words published by this user. The dimensionality of user embedding vector and word embedding vector are both set to be 200.

Model Training and Parameter Settings: In the classification (or regression) phase, we employ L2-regularized SVM (or support vector regression, SVR) model in the LibLinear package[6]. For Word Embeddings and User Embeddings, the mini-batch size of the stochastic gradient descent is set to be 1; the learning rate is set to be $\rho_t = \rho_0(1 - t/T)$, in which T is the total number of edge samples and $\rho_0 = 0.025$; the number of negative samples is set to be 5; the window size is set as 5 in Skip-gram.

Evaluation Metrics and Significance test: For gender classification task, the performance is evaluated using Macro-F1 (F). For age regression task, we employ the coefficient of determination R^2 to measure the regression performance (Cameron, 1996). Furthermore, t-test is used to evaluate the significance of the performance difference between two approaches (Yang and Liu, 1999).

4.2 Impact of Using User-pivotal_word Network G_{uv}^{p}

In order to test the effectiveness of using user-pivotal_word network G_{uv}^{p} to form the cross-media user-user network, the following two approaches are implemented for cross-media user profiling.

- **Cross-uu without Pivotal Words:** our model with **User Embeddings** as input. The user embeddings are learned from the cross-media user-user network which is composed of two user-user networks, i.e., $G_{uu}^{cross} = G_{uu}^{s} + G_{uu}^{t}$.

- **Cross-uu with Pivotal Words:** our model with **User Embeddings** as input. The user embeddings are learned from the cross-media user-user network which is composed of two user-user networks and one user-pivotal_word network, i.e., $G_{uu}^{cross} = G_{uu}^{s} + G_{uu}^{t} + G_{uv}^{p}$.

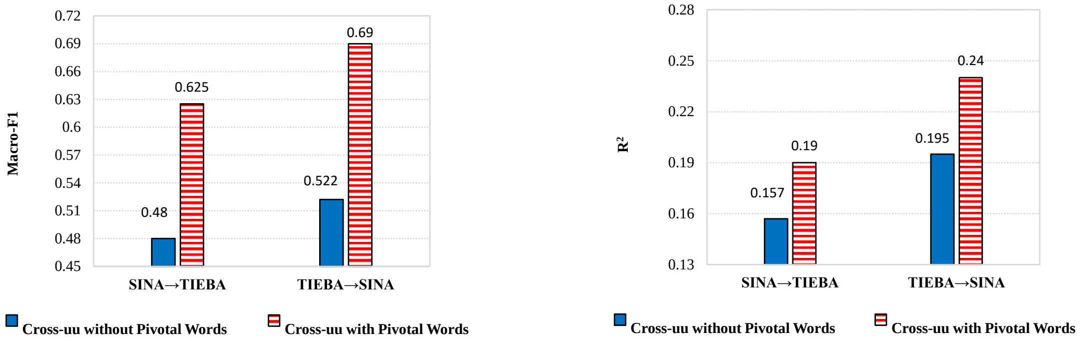
Figure 7(a) and 7(b) shows the results of the two approaches to cross-media gender classification and cross-media age regression respectively. From this figure, we can see that **Cross-uu without Pivotal Words** performs similar to a random classification, achieving about 50% in terms of Macro-F1 in gender classification task. This result is not strange because the involved two user-user networks share nothing and thus could not capture the relationship among the users in the two social media. In contrast, **Cross-uu with Pivotal Words** performs much better than a random performance. Especially, in gender classification task, in the case of TIEBA → SINA, our approach achieves 0.69 in terms of Macro-F1,

[3] `http://www.ictclas.org/ictclas\_download.aspx`

[4] `http://nlp.stanford.edu/software/lex-parser.shtml`

[5] `https://github.com/dav/word2vec`

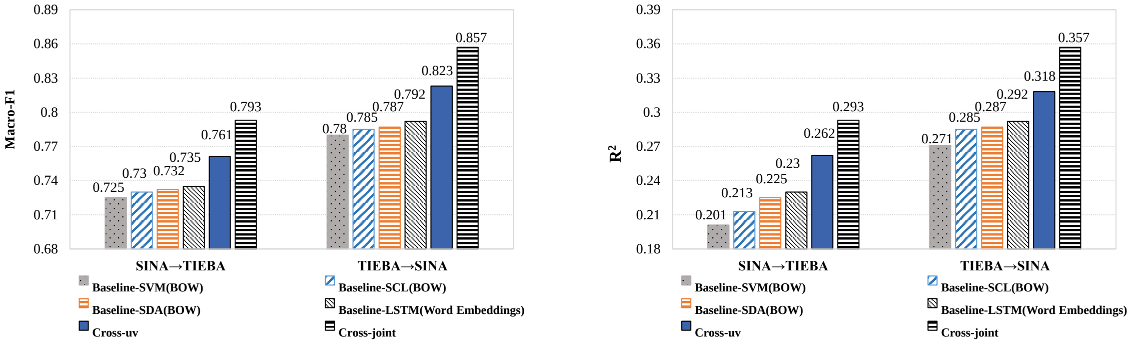
[6] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`



(a) Cross-media gender classification

(b) Cross-media age regression

Figure 7: Performances of two approaches (i.e., using or not using user-pivotal_word network).



(a) Cross-media gender classification

(b) Cross-media age regression

Figure 8: Performances of different approaches.

which is 16.8% better than **Cross-uu without Pivotal Words**. Moreover, in age aggression task, **Cross-uu with Pivotal Words** still outperforms **Cross-uu without Pivotal Words** about 4.5% in terms of R^2. These results confirm that constructing the user-pivotal_word network is beneficial for learning a better user embedding for cross-media user profiling.

4.3 Performance Comparison

For thorough comparison, several approaches are implemented for cross-media user profiling:

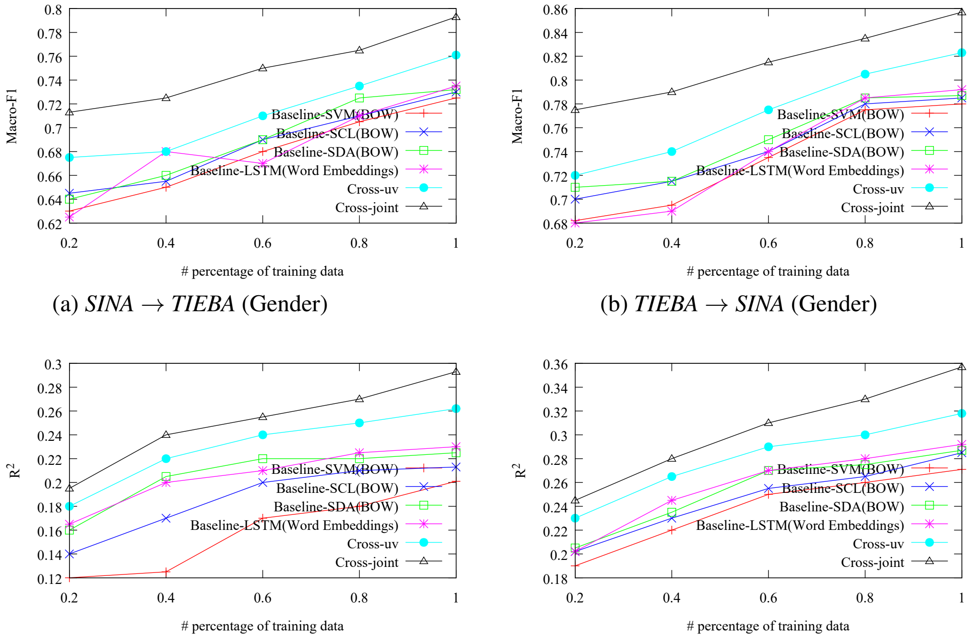
- **Baseline-SVM(BOW):** a SVM classifier (or SVR regressor) trained with only labeled data in the *source* media and the representation model is **BOW**. This approach is proposed by Li et al. (2015).

- **Baseline-SCL(BOW):** a famous textual domain adaptation approach named SCL which bridges the knowledge between the *source* and *target* domains using some pivotal features (Blitzer et al., 2007) and the representation model is **BOW**. We consider the *source* media as the *source* domain and the *target* media as the *target* domain. The number of the pivotal features is set to be 200.

- **Baseline-SDA(BOW):** another famous textual domain adaptation approach with stacked denoising auto-encoder to extract meaningful representation (Glorot et al., 2011) and the representation model is **BOW**. Similar to SCL, we consider the *source* media as the *source* domain and the *target* media as the *target* domain.

- **Baseline-LSTM(Word Embeddings):** a LSTM classification (or regression) model using **Word Embeddings** as input. This is a state-of-the-art approach to user profiling proposed by Wang et al. (2017).

- **Cross-uv:** our model with **User Embeddings** as input. The user embeddings are learned from the cross-media user-word network, i.e., G_{uv}^{cross}.



(a) *SINA* → *TIEBA* (Gender)

(b) *TIEBA* → *SINA* (Gender)

(c) *SINA* → *TIEBA* (Age)

(d) *TIEBA* → *SINA* (Age)

Figure 9: Performance of different approaches on various sizes of training data

- **Cross-joint:** our model with **User Embeddings** as input. The user embeddings are learned from the heterogeneous user network, i.e., G_{joint}.

Figure 8(a) and 8(b) shows the comparison results of different approaches to cross-media gender classification and cross-media age regression respectively. From these two figures, we can see that: The textual domain adaptation approaches including **Baseline-SCL(BOW)** and **Baseline-SDA(BOW)** are effective for cross-media user profiling and they both performs slightly better than **Baseline-SVM(BOW)**. Our approach **Cross-uv** performs consistently better than the textual domain adaptation approaches. The improvements are about 2% in terms of both Macro-F1 and R^2. The significance test shows that the improvements are significant (p-value$<$ 0.05). Besides, our approach performs better than **Baseline-LSTM(Word Embeddings)**, which indicates that our approach yields better user embeddings than the traditional word embedding approach.

Our approach **Cross-joint** performs best among all approaches. Compared to **Baseline-SVM(BOW)**, the average improvement is impressive, reaching 7.25% (Macro-F1) and 8.9% (R^2) in gender classification and age regression respectively. Furthermore, **Cross-joint** significantly performs better than **Cross-uv** (p-value$<$ 0.05), which encourages to incorporate the social link information for cross-media user profiling.

Figure 9 shows the performance of the approaches to cross-media user profiling by changing the sizes of the training data in the *source* media. From this figure, we can see that when the size of training data is small, the performance of **Baseline-LSTM(Word Embeddings)** is rather unstable, performing worse than **Baseline-SDA(BOW)** and **Baseline-SCL(BOW)** in some times and could even performs worse than **Baseline-SVM(BOW)**. However, our approach **Cross-joint** and **Cross-uv** consistently perform better than the other approaches, whatever sizes of training data are used.

5 Conclusion

In this paper, we propose a user embedding learning method to address cross-media user profiling. Specifically, we first propose a heterogeneous user network composed of two cross-media networks for learning user embedding and then employ the user embedding to train the model for classification (or regression). Empirical studies demonstrate that our approach significantly outperforms other strong baseline approaches in both cross-media gender classification and cross-media age regression tasks.

In our future work, we would like to incorporate some other types of information, e.g., label information, to better learn user embedding. Moreover, we would like to apply our approach to some other user profiling tasks, e.g., user profession identification.

Acknowledgments

We thank anonymous reviewers for their valuable suggestions and comments. The research work is supported by three National Natural Science Foundation of China (No.61751206, No.61331011 and No.61672366) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL-2007*.

A. Colin Cameron. 1996. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2):209–220.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of EMNLP-2013*, pages 1136–1145.

Malcolm Corney, Olivier Y. de Vel, Alison Anderson, and George M. Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Proceedings of ACSAC-2002*, pages 282–289.

Philip Gianfortoni, David Adamson, and Carolyn P Rosé. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of EMNLP-2011*, pages 49–59. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of ICML-2011*, pages 513–520.

Shoushan Li, Jingjing Wang, Guodong Zhou, and Hanxiao Shi. 2015. Interactive gender inference with integer linear programming. In *Proceedings of IJCAI-2015*, pages 2341–2347.

Shoushan Li, Bin Dai, Zhengxian Gong, and Guodong Zhou. 2016. Semi-supervised gender classification with joint textual and social modeling. In *Proceedings of COLING-2016*, pages 2092–2100.

James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. 2014. Age and gender identification in social media. In *Proceedings of CLEF-2014*, pages 1129–1136.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Saif M. Mohammad and Tony Yang. 2013. Tracking sentiment in mail: How genders differ on emotional axes. *CoRR*, abs/1309.6347.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of EMNLP-2010*, pages 207–217.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of ICWSM-2010*.

Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 750–784.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of CIKM-2011*, pages 37–44.

Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *Proceedings of ACL-2015*, pages 1754–1764.

Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of NIPS-2011*, pages 693–701.

Maarten Sap, Gregory J. Park, Johannes C. Eichstaedt, Margaret L. Kern, David Stillwell, Michal Kosinski, Lyle H. Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of EMNLP-2014*, pages 1146–1151.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of CONLL-2011*, pages 78–86.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI-2006*, pages 199–205.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015a. PTE: predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of SIGKDD-2015*, pages 1165–1174.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015b. LINE: large-scale information network embedding. In *Proceedings of WWW-2015*, pages 1067–1077.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP-2013*, pages 1815–1827.

Jingjing Wang, Shoushan Li, and Guodong Zhou. 2017. Joint learning on relevant user attributes in micro-blog. In *Proceedings of IJCAI-2017*, pages 4130–4136.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-1999*, pages 42–49.

Dong Zhang, Shoushan Li, Hongling Wang, and Guodong Zhou. 2016. User classification with multiple textual perspectives. In *Proceedings of COLING-2016*, pages 2112–2121.