Improving Users' Demographic Prediction via the Videos They Talk about

Yuan Wang, Yang Xiao, Chao Ma, and Zhen Xiao
Department of Computer Science, Peking University, Beijing 100871, China
{wangyuan, xiaoyang, machao, xiaozhen}@net.pku.edu.cn

Abstract

In this paper, we improve microblog users' demographic prediction by fully utilizing their video related behaviors. First, we collect the describing words of currently popular videos, including video names, actor names and video keywords, from video websites. Secondly, we search these describing words in users' microblogs, and build the direct relationships between users and the appeared words. After that, to make the sparse relationship denser, we propose a Bayesian method to calculate the probability of connections between users and other video describing words. Lastly, we build two models to predict users' demographics with the obtained direct and indirect relationships. Based on a large real-world dataset, experiment results show that our method can significantly improve these words' demographic predictive ability.

1 Introduction

Recent studies have indicated that users' demographics can be predicted from their linguistic characteristics. A typical practice is cutting the text into a bag of words and training a linear classifier. Although this practice can achieve an acceptable result in simple tasks such as predicting gender and age, it loses some important information about the text structure and does not fully use the relationship between words.

Nowadays, people spend a lot of time on videos and social media which provide them with access to post views and comments. Weibo is one of the biggest microblogging platforms in China. More than one third of the "Weibo Trends"[1] are about videos. Generally, people with different demographic attributes usually have different tastes for videos (Abisheva et al., 2014). For example, in China people who watch English drama tend to be well-educated. Here is a question: if the video related information in users' weibo messages can be fully used, will the users' demographic prediction be improved?

One challenge is that many users do not directly mention the video names in their weibo messages. Instead, they make comments on the actors or the plots. If a person likes "Big Bang Theory", he may post "*Will the Big Bang Theory last into the next century?*" where the sitcom's name is mentioned directly, or "*Sheldon is so cool, I love him!*" which talks about an actor of the sitcom. Both posts indicate the user is interested in "Big Bang Theory". When involving the demographic prediction, however, the traditional "bag of words based" model cannot extract the above information effectively. Some previous works use topic models such as LIWC (Pennebaker et al., 2001) or LDA (Blei et al., 2003) to detect the relations among users' words. Usually, they suffer from the short length of weibo messages and the number of topics. In addition, the lifespan of most popular video programs is not very long, which renders traditional topic models inefficient.

Fortunately, there exist some third-party video websites, such as *youtube.com* and *youku.com*, from which we can get the most popular videos. For each video, there is usually a homepage with a actor list

[1] http://d.weibo.com/100803

and also a comments section, and we can calculate the video's Top TF-IDF words (keywords) based on these comments. Here we define the *video name*, *actor name* and *keyword* to be three different kinds of "video describing words". The relationships among these words can be used to better understand weibo users' video related behaviors. This approach can be applied to other kinds of words, such as describing words on books and music. This paper focuses on the video as an example.

After obtaining the video describing words, we build three matrices to represent the direct and indirect relationships between weibo users and these words. They are User-Video Matrix, User-Actor Matrix and User-Keyword Matrix, respectively. At beginning, these three matrices are sparse because they only represent the direct relationships, which means that only when the words appear in user's weibos, the corresponding position will be set. After that, we propose a "hidden layer" to detect the indirect relationships, making them denser.

With these indirect relationships, we can improve users' demographic predictions, including *gender*, *age*, *education background*, and *marital status*. This paper makes the followings three contributions:

1. By construct three matrices, we detect the direct and indirect relationships between weibo users and video describing words.

2. Two models are proposed to predict users' demographics by using both direct and indirect relationships.

3. Experiment results prove that our efforts can significantly improve the predictive accuracy, compared with the existing research.

The rest of this paper is organized as follows. Section 2 introduces the dataset and demographics. Section 3 introduces how to make full use of video related behaviors. Section 4 presents experimental results. Finally, we review related work in Section 5, and draw conclusions in Section 6.

2 Dataset and Demographics

2.1 Dataset

We collected 2,970,642 microblog users from Weibo (http://weibo.com), the largest microblog service in China, as our dataset. To avoid spam users (sometimes called robot users), we only collected *verified users* and *users followed by verified user*. Weibo conducts manual verifications to make sure the verified users provide real and authentic profile information. Table 1 presents four target demographic attributes and the completion rates (ratio of effective users). All data is either through Open API or publicly available. No private data is used in the experiment.

We also collected 847 popular video programs from YISQ (4 popular video websites in China: *youku*, *iqiyi*, *sohu*, *qq*). These videos mainly fall into three types: movie, tv play, and variety shows. We downloaded these videos' Homepages and extracted their actors and TOP20 TF-IDF words. The statistics are shown in Table 2.

2.2 Ground Truth

One problem of our dataset is it contains celebrities, while our model mainly targets ordinary weibo users. We implement a filter to exclude celebrities based on their large numbers of followers (>50000 as default), making the ground truth more representative. Besides, users with less than 100 messages are discarded. At last, we obtain 742,323 accounts with both their demographics and messages.

2.3 Demographics

As Table 1 shows, the demographic attributes concerned in this paper include gender, age, education background, and marital status:

Gender (Binary): the gender prediction is a typical binary classification task: male, female.

Age (4-Class): because there is only a handful of(<1%) user older than 45, we classify users into the following four age groups: Teenage (<18), Youngster (18-24), Young (25-34), Mid-age (>34).

Education Background (Binary): we categorize users' education background into two groups: university, non-university.

Marital Status (Binary): marital status is also simplified to a binary classification task: single, non-single.

3 Our Model

In this section, we introduce the framework, which contains four steps.

Attribute	Completion Rate	Categories
Gender	95.019%	Male, Female
Age	18.604%	Teenage (<18), Youngster (18-24), Young (25-34), Mid-age(>34)
Education BG	17.443%	University, Non-University
Marital Status	2.203%	Single, Non-Single

Table 1: Demographic attributes and corresponding categories

	Video	Actor	Keyword
Variety show	344	1007	2925
Movie	306	741	2049
TV	197	515	1302
Total	**847**	1422	4094

Table 2: Statics of video relevant information (There is an overlap between the three collections of actors and keywords.)

The first step generates the "Video describing words" and represents user as two vectors (V_v, V_o). V_v consists of user's "video describing words" and V_o consists of user's "other words". At first, V_v only contains user's direct relationships.

> *V_v: video describing words (direct)*
> *V_o: other words*
> *V_a: V_v + V_o*

The second step detects the indirect relationships between users and videos. For example, if a user mentioned "Robert Downey Jr", we believe he has an indirect relationships with "Iron Man" movie. By doing so, we add user's indirect relationships into his V_v, getting a denser vector V_v'.

> *V_v': video describing words (direct+indirect)*
> *V_a': V_v' + V_o*

The third step proposes two models respectively to evaluate whether those indirect relationships, discovered in second step, can be used to develop a more accurate prediction model.

The fourth step represents weibo user with the combination of V_v' and V_o, and use the combination to train a linear SVM to evaluate whether this effort can make the prediction better.

3.1 Discover Indirect Relationships

If a user mentioned a video's name directly, we believe there is a direct relationship between them. The rests are unobvious relationships. In this part, we calculate whether these unobvious relationships can be transformed into indirect ones.

3.1.1 User-Video Matrix

Firstly, we detect whether a user directly mentioned a video program in his weibo messages. There are two scenarios: the first is this user posts a message containing the video's name directly, and the other is this user reposts a message containing the video's name. In this paper, we believe these two scenarios both indicate there is a direct relationship between the user and the video, and do not make a distinction between them. Till now, we construct a Direct User-Video Matrix (DUVM) to denote all the direct relationships between users and videos.

Step 1: We know each video program v_n contains some actors a_{nj} and keywords w_{ni}. We can calculate $P(v_n)$, $P(a_{nj}|v_n)$ and $P(w_{ni}|v_n)$ in Step 1. $P(v_n)$ represents the probability that a person has watched the n_{th} video. $P(w_{ni}|v_n)$ represents the probability that a person, who has watched the n_{th} video, mention the ni_{th} keyword. $P(a_{nj}|v_n)$ is the probability that a person, who has watched the n_{th} video, mention the nj_{th} actor.

$P(v_n)$ = num (users watched the n_{th} video) / num (users)

$P(w_{ni}|v_n)$ = num (users watched the n_{th} video and mentioned the ni_{th} keyword) / num (users watched the n_{th} video)

$P(a_{nj}|v_n)$ = num (users watched the n_{th} video and mentioned the nj_{th} actor) / num (users watched the n_{th} video)

Step 2: In step 2, If a user doesn't mention a video's name directly, but mentions the video's related actors (A_k) and keywords (W_m), we can update his unobvious user-video relationships according to a Bayesian framework.

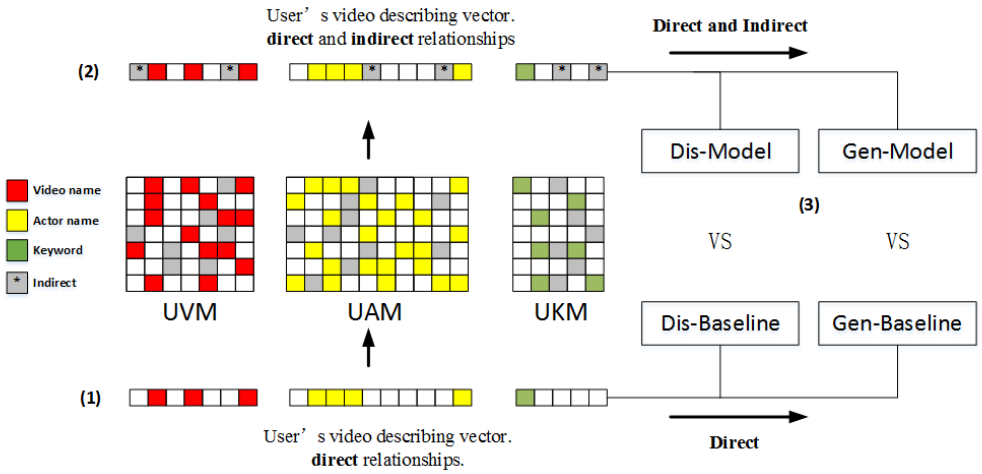


Figure 1: (1) At first, identify the describing words from users microblogs, which builds the direct relationships between users and these words. (2) By construct three matrices, we detect the indirect relationships between weibo users and video describing words. (3)Two models are proposed to predict users demographics by using both direct and indirect relationships.

$$P(v_n|W_m, A_k) = \frac{P(W_m, A_k|v_n) * P(v_n)}{P(W_m, A_k)} = \frac{\prod_{w_{ni} \in W_m} P(w_{ni}|v_n) * \prod_{a_{nj} \in A_k} P(a_{nj}|v_n) * P(v_n)}{P(W_m, A_k)} \quad (1)$$

Through Step 2, we can discover some new indirect relationships and update UVM. Go back to Step 1 and iterate until converges, we can get the Final UVM at last.

3.1.2 User-Actor Matrix

Every video program has several actors, and the relationships between weibo users and actors may contribute to the demographic prediction either. So we build the UAM, where each row represents a weibo user and each column represents an actor.

There are two case that the element of UAM will be set to true: (1) the user 'i' directly mentioned actor 'j' in his weibo messages (including post and repost); (2) the user 'i' has watched video 'v', and actor 'j' participate in video 'v'. The second case needs UVM's help. We suppose these two cases affect the value equally in this paper.

3.1.3 User-Keyword Matrix

We can find several keywords to describe each video from their Homepages. For instance, we get "Paul Walker", "fight", and "car" to describe "Furious 7".

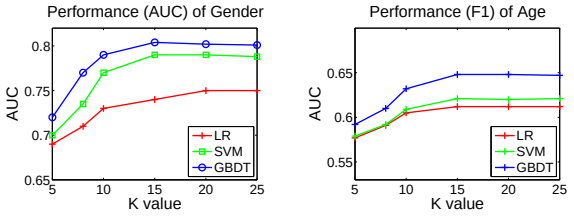
Each row of UKM represents a weibo user and each column represents a keyword of a certain video. (1) If we find a user has watched the "Furious 7", no matter direct or indirect relationship, we can set the columns of user's "Furious 7" keywords to true. (2) The value can be set to true either if the user directly mentioned these keywords.

3.2 Two Indirect Relationship Based Models

In this part, two models are proposed to predict users' demographics by using both direct and indirect relationships.

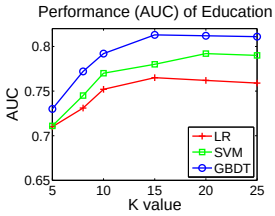
3.2.1 Discriminant Model (Dis-Model)

Given three matrices, the intuitive way to predict users' demographics is using Collaborative Filtering. However, finding the similar users directly based on the vector similarity is not a good idea, because a substantial part of users have ever watched no more than 10 videos. Matrix Factorization has been proven useful to address data sparsity, for the reduced orthogonal dimensions are less noisy than the original data and can capture the latent associations between users and videos. In our Dis-Model, we utilize the factorization machines (Rendle, 2010) to deal with UVM, UAM, and UKM, reducing the length of user's dimensionality from videos' number (actors' number, keywords'

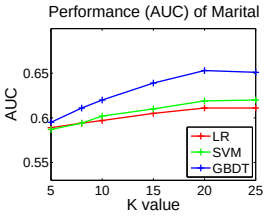


(a) AUC of Gender

(b) F1 of Age



(c) AUC of Education BG



(d) AUC of Marital Status

Figure 2: Performance of different classifiers (LR, SVM, GBDT) for Dis-Model with varying K.

number) to a smaller value K. Every weibo user can be represented by the combination of these three K-length vectors.

Over the last several decades, many kinds of discriminant classifier have been created. For our four tasks, we compared Logistic Regression (LR), Support Vector Machine (SVM), and Gradient Boosted Decision Tree (GBDT). Figure 2 illustrates their performance, where GBDT performs the best in all K values. When K increases from 5 to 20, all classifiers' results are all getting better and tend to be stable when K is bigger than 20. So we choose GBDT as our default base classifier and K=20 as default value.

3.2.2 Generative Model (Gen-Model)

We start with introducing an important concept: video demographic tendency, which means to what extent a video belongs to a specified demographic group. For example, if 90% audiences of a movie are males, we define its demographic tendency on male as 90%. The actor tendency and keyword tendency can be calculated in the same way.

In the Gen-Model, (1) we firstly calculate each video's (actor, keyword) demographic tendency according to its audiences (known demographics). (2) Based on the demographic tendency of videos (actors, keywords), we predict user's (unknown) demographics via a Bayesian method. (3) At last, we propose a smooth step to adjust the result.

(1) Calculate video demographic tendency

At first, we calculate every video demographic tendency as Equation 2:

$$p(c|v_j) = \frac{\sum_{i=1}^{n}(r_{ij} * u_i(c))}{\sum_{i=1}^{n} r_{ij}} \tag{2}$$

$P(c|v_j)$ represents the j_{th} video's demographic tendency on c, where c is the demographic attribute. r_{ij} will be set to 1 if the i_{th} user has watched the j_{th} video, otherwise set to 0. $u_i(c)$ is a boolean, representing whether the i_{th} user has the attribute c.

(2) Calculate user demographic attribute

In this step, we predict users' demographics according to the demographic tendency of the videos they has watched. Suppose user's viewing habits are independent, we can calculate the probability of $P(c|u_i)$ as Equation 3:

$$\begin{aligned} P(c|u_i) &\propto P(c|\{V\}) \\ &\propto P(\{V\}|c) * P(c) \\ &\propto \prod_{v_j \in \{V\}} P(v_j|c) * P(c) \\ &= \frac{\prod_{v_j \in \{V\}} P(c|v_j) * P(v_j)}{P(c)} * P(c) \\ &\propto \prod_{v_j \in \{V\}} P(c|v_j) \end{aligned} \tag{3}$$

$\{V\}$ represents the collection of videos watched by u_i. $P(c|v_j)$ is the j_{th} video's demographic tendency on c, as the previous part described.

(3) Smooth the result

Based on the fact that people in same demographic group may have similar behaviors, we deploy a smooth component to adjust the value of $P(c|v_j)$ and $P(c|u_i)$ according to their top *n* neighbors. As mentioned above, we use factorization machines to transform the user and video vectors into low-dimensional (K=20) ones. The distance is calculated by Euclidean Distance. The video, actor, and word have the same treating process, so we introduce the video as representative.

Smooth the Video's Demographic Tendency: Base on video v_j's top n neighbors, we can calculate its neighbors' average demographic tendency $P(c|nbr(v_j))$, where $P(c|v_{nbj})$ is v_j's nbj_{th} neighbor's demographic tendency.

$$p(c|nbr(v_j)) = \frac{\sum_{j=1}^{n} P(c|v_{nbj})}{n} \quad (4)$$

Therefore, we can smooth v_j's demographic tendency by:

$$P(c|v_j) = \alpha * P(c|v_j) + (1-\alpha) * P(c|nbr(v_j)) \quad (5)$$

α is the parameter to control the top n neighbors' influence. In this paper, we compared ten values of α and chose 0.7 as default. With the same process, n is set to 10 as default.

Smooth the User's Demographic Result: The user side smooth procedure is similar to the video side, except user's $P(c|nbr(u_i))$ is affected by three kinds of neighbors (u_{nbvi}, u_{nbai}, u_{nbwi}).

$$p(c|nbr(u_i)) = \frac{\sum_{i=1}^{n} P(c|u_{nbvi})}{3n} + \frac{\sum_{i=1}^{n} P(c|u_{nbai})}{3n} + \frac{\sum_{i=1}^{n} P(c|u_{nbwi})}{3n} \quad (6)$$

Just like video's smooth process, we adjust u_i's demographic attributes by:

$$P(c|u_i) = \alpha * P(c|u_i) + (1-\alpha) * P(c|nbr(u_i)) \quad (7)$$

The smooth component is deployed as an iterative procedure, and keeps running until each $P(c|u_i)$ became stable.

Two Baselines: To validate whether those indirect relationships can improve the predictions, we build two baseline models: Dis-Baseline and Gen-Baseline. While our two models use the V_v' as input, these two baseline models use the raw V_v. These two baseline models adopt the same architecture with our proposed two models. The only difference is the input data.

3.3 Fusion Model

As described above, we discovered the indirect relationships between users and video describing words, and demonstrated this effort can leading a better result than directly train the classifier.

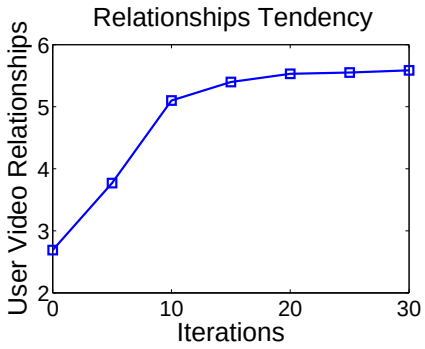


Figure 3: Tendency of User-Video relationship number.

But pre-existing models commonly utilize all the words in user's weibo messages. So we need to find out whether our hard-earned improvement would be submerged by those "Non video describing words". We train a Fusion Model using all the words in weibo messages and indirect relationships together, and compare it with a baseline model, who only use all the words (without indirect relationships).

Fusion Baseline: Many pre-existing methods (Burger et al., 2011; Tu et al., 2015) chose linear model as their text classifier, for linear model is suitable for text categorization tasks. We choose L1-regularized linear SVM as our Fusion Model and Fusion-Baseline's classifier. The only difference between them is the input data (V_v' + V_o **vs** V_v + V_o).

4 Experiment Results

We conducted a 10-fold cross validation to demonstrate our framework's effectiveness, where 8 parts for training, 1 parts for validation and 1 parts for testing by default. The performance of presented methods were evaluated using the Precision, Recall and Macro-F1 measures. Binary classification tasks were also measured by Area Under the ROC Curve (AUC).

4.1 Indirect Relationships Evaluation

In our dataset, each user directly mention 2.6 video programs on average and only 0.7% has more than 10 direct relationships. As shown in Figure 3, more and more indirect relationships arise along with the iterations. User's relationship number (direct + indirect) stabilized at 5.7 on average and 13% of them is bigger than 10.

To answer whether these indirect relationships

		Precision	Recall	F1	AUC
Gender	Dis-Baseline	0.720	0.714	0.717	0.730
	Dis-Model	0.786	0.779	0.783	**0.812** ↑ 11.2%
	Gen-Baseline	0.701	0.687	0.694	0.707
	Gen-Model	0.799	0.802	0.801	**0.825** ↑ 16.7%
Age	Dis-Baseline	0.569	0.541	0.554	*
	Dis-Model	0.642	0.653	**0.648** ↑ 16.8%	*
	Gen-Baseline	0.529	0.504	0.516	*
	Gen-Model	0.663	0.645	**0.654** ↑ 26.7%	*
Education BG	Dis-Baseline	0.707	0.716	0.711	0.730
	Dis-Model	0.788	0.801	0.795	**0.809** ↑ 11.1%
	Gen-Baseline	0.680	0.659	0.669	0.690
	Gen-Model	0.790	0.808	0.799	**0.812** ↑ 17.7%
Marital Status	Dis-Baseline	0.565	0.549	0.557	0.571
	Dis-Model	0.657	0.640	0.648	**0.659** ↑ 15.4%
	Gen-Baseline	0.572	0.550	0.560	0.581
	Gen-Model	0.682	0.691	0.687	**0.696** ↑ 19.8%

Table 3: Prediction accuracy based on users' video describing words. Classes have been balanced.

can make the prediction better, we compared our two models (Dis-Model & Gen-Model) with two baseline models. We also compared their performance on different user groups categorized by user-video relationship number.

Gender: As Table 3 shows, our two models both have a significant improvement compared to the baseline models. The Gen-Model achieve the best performance (AUC 0.825) in terms of all the measurement. As Figure 4(a) shows, with the number growth, our two models' AUC scores are both getting better. Surprisingly, when the number is bigger than 10, the Gen-Model even get a similar performance of the model using all of the user's words.

Age: In the age task, our two models both outperformed the baseline models significantly, and the generative model performs better (F1 0.654) too. We analyzed the result and found the "youngster" and "young" share the similar watching habits in Weibo. It's hard to pick out a 23 years old user from the 28 years old group. As Figure 4(b) shows, our two models' F1 scores are both getting better along with the growth of user-video relationship number.

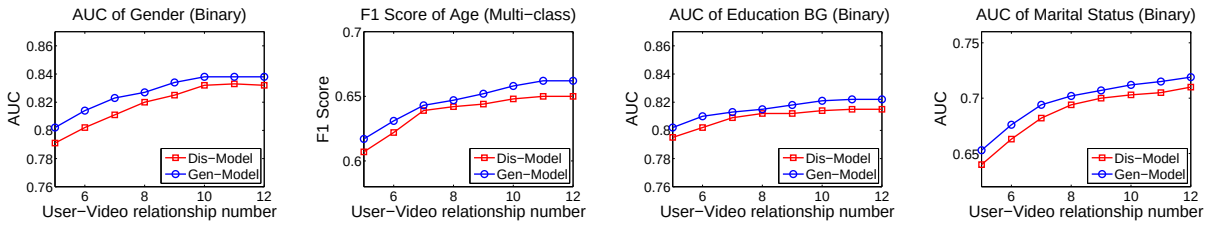
Education Background: Not surprisingly, our two models obviously outperform the result over two baseline models. This result indicates that people in different education background has visible different tastes on video programs.

Marital Status: Table 3 presents the results of marital status. We notice that the performance of our two model is still reasonable, but is worse than gender and education tasks. In addition to that this task is more difficulty, another reason is when a user gets married, he might not update the information in his online profile.

Remark: Experiment results show that our method can significantly improve these words' demographic predictive ability by more than 15% on average. 10 videos is good enough to portray a weibo user, and can achieve reasonable results in these 4 inference tasks. The video related behavior is efficient on predicting gender and education, for people on these two tasks have visible different inclinations. Inferring age and marital status is not easy, but our two models still achieve reasonable improvements. In general, our two models both get significantly better results than baselines. The Gen-Model is a better choice by contrast.

4.2 Fusion Model Evaluation

After we obtained the potential predictive ability of indirect relationships, we also need to find out whether it can help pre-existing model perform



(a) AUC of Gender (b) F1 Score of Age (c) AUC of Education BG (d) AUC of Marital Status

Figure 4: Prediction result with varying User-Video relationship numbers.

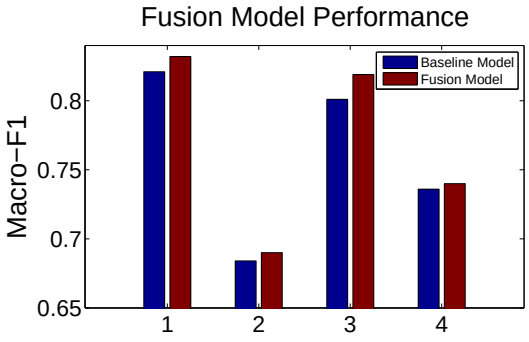


Figure 5: Results of Fusion Model evaluation (Macro-F1).

better. We compare the Fusion Baseline (V_v+V_o) with our Fusion Model (V_v'+V_o). As Figure 5 shows, Fusion Model's performance is better than Fusion Baseline's in all four tasks. The improvement is about 2-3% on average. As above mentioned, our approach can be applied to other kinds of words, such as describing words on books and music. So there is some room for improvement.

5 Related work

In this section, we briefly review the research works related to our work.

Many researches (Kumar and Tomkins, 2010; Goel et al., 2012) found users belong to different demographic groups behave differently. (Hu et al., 2007; Murray and Durrell, 2000; Goel et al., 2012; Kosinski et al., 2012) showed that age, gender, education level, and even personality can be predicted from people's webpage browsing logs. (Kosinski et al., 2013; Schwartz et al., 2013; Youyou et al., 2015) showed computers' judgments of people's personalities based on their Facebook Likes are more accurate and valid than judgments made by their close acquaintances. (Malmi and Weber, 2016) showed users' demographics also can be predicted based on theirs apps. Apart from the browsing behaviors, there also exist some works based on user's linguistic characteristics. (Schler et al., 2006) analyzed tens of thousands of blogs and indicated significant differences in writing style and word usage between different gender and age groups. The similar result also showed in (Luyckx and Daelemans, 1998; Oberlander and Nowson, 2006; Mairesse et al., 2007; Nowson, 2007; Gill et al., 2009; Rosenthal and McKeown, 2011). There are some works (Bi et al., 2013; Weber and Jaimes, 2011; Weber and Castillo, 2010) on predicting search engine user's demographics based on their search queries. (Hovy, 2015) investigated the influence of user's demographics on better understanding their online reviews. (Otterbacher, 2010) used logistic regression model to infer users gender based on the content of movie reviews.

Many researches focused on the twitter users. In the Author Profiling task at PAN 2015 (Rangel et al., 2015), participants approached the task of identifying age, gender and personality traits from Twitter. (Nguyen et al., 2013) explored users' age prediction task based on their tweets, achieving better performance than humans. (Burger et al., 2011) studied the gender predictive ability of twitter linguistic characteristics, reached 92% accuracy. (Pennacchiotti and Popescu, 2011) proposed a GB-DT model to predict users' age, gender, political orientation and ethnicity by leveraging their observable information. (Culotta et al., 2015) predicted the demographics of Twitter users based on whom they follow, and (Zhong et al., 2015) predicted the microblog user's demographic attributes only by their chick-ins. In (Li et al., 2014), job and education attributes are extracted by combining a rule based approach with a probabilistic system. There are also some works based on users' social relationships

(Mislove et al., 2010; Henderson et al., 2012; Zhao et al., 2013).

6 Conclusion

Our motivation on writing this paper is user's video related behavior is usually under-utilized on demographic prediction tasks. With the help of third-party video sites, we detect the direct and calculate the indirect relationships between users and video describing words, and demonstrate this effort can improve the accuracy of users' demographic predictions. To our knowledge, this is the first work which explores demographic prediction by fully using users' video describing words. This framework has good scalability and can be applied on other concrete features, such as user's book reading behaviors and music listening behaviors.

Acknowledgments

This work was supported by the National Grand Fundamental Research 973 Program of China under Grant No.2014CB340405 and the National Natural Science Foundation of China under Grant No.61572044.

References

Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. 2014. Who watches (and shares) what on youtube? and when?: using twitter to understand youtube viewership. In *Proceedings of WSDM*, pages 593–602. ACM.

Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of WWW*, pages 131–140. International World Wide Web Conferences Steering Committee.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the EMNLP*, pages 1301–1309. Association for Computational Linguistics.

Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. 2015. Predicting the demographics of twitter users from website traffic data. In *Proceedings of AAAI*, pages 72–78.

Alastair J Gill, Scott Nowson, and Jon Oberlander. 2009. What are they blogging about? personality, topic and motivation in blogs. In *Proceedings of ICWSM*.

Sharad Goel, Jake M Hofman, and M Irmak Sirer. 2012. Who does what on the web: A large-scale study of browsing behavior. In *Proceedings of ICWSM*.

Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. 2012. Rolx: structural role extraction & mining in large graphs. In *Proceedings of SIGKDD*, pages 1231–1239. ACM.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*.

Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of WWW*, pages 151–160. ACM.

Michal Kosinski, David Stillwell, Pushmeet Kohli, Y-oram Bachrach, and Thore Graepel. 2012. Personality and website choice.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Ravi Kumar and Andrew Tomkins. 2010. A characterization of online browsing behavior. In *Proceedings of WWW*, pages 561–570. ACM.

Jiwei Li, Alan Ritter, and Eduard H Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of ACL*, pages 165–174.

Kim Luyckx and Walter Daelemans. 1998. Using syntactic features to predict author personality from text. *Science*, 22:319–346.

Francois Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, pages 457–500.

Eric Malmi and Ingmar Weber. 2016. You are what apps you use: Demographic prediction based on user's apps. *arXiv preprint arXiv:1603.00059*.

Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. 2010. You are who you know: inferring user profiles in online social networks. In *Proceedings of WSDM*, pages 251–260. ACM.

Dan Murray and Kevan Durrell. 2000. Inferring demographic attributes of anonymous internet users. In *Web Usage Analysis and User Profiling*, pages 7–20. Springer.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "how old do you think i am?"; a study of language and age in twitter. In *Proceedings of ICWSM*. AAAI Press.

Scott Nowson. 2007. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proceedings of ICWSM*. Citeseer.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics.

Jahna Otterbacher. 2010. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of CIKM*, pages 369–378. ACM.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *Proceedings of ICWSM*, pages 281–288.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September.

Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of ACL*, pages 763–772. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Cunchao Tu, Zhiyuan Liu, and Maosong Sun, 2015. *Social Media Processing: 4th National Conference, SMP 2015, Guangzhou, China, November 16-17, 2015, Proceedings*, chapter PRISM: Profession Identification in Social Media with Personal Information and Community Structure, pages 15–27. Springer Singapore, Singapore.

Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *Proceedings of SIGIR*, pages 523–530. ACM.

Ingmar Weber and Alejandro Jaimes. 2011. Who uses web search for what: and how. In *Proceedings of WSDM*, pages 15–24. ACM.

Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

Yuchen Zhao, Guan Wang, Philip S Yu, Shaobo Liu, and Simon Zhang. 2013. Inferring social roles and statuses in social networks. In *Proceedings of SIGKDD*, pages 695–703. ACM.

Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location checkins. In *Proceedings of WSDM*, pages 295–304. ACM.