PROCEEDINGS OF SCIENCE

A New Method for Constructing TV User Profiles Based on Micro-blog

Jiashou Chen[1]
Beijing University of Posts and Telecommunications,
Beijing, 100876, China
E-mail: chenjiashou001@163.com

Bo Lu
China Academy of Information and Communications Technology,
Beijing, 100191, China
E-mail: lubo@caict.ac.cn

Zhiqian Wang
Beijing University of Posts and Telecommunications,
Beijing, 100876, China
E-mail: wzq@bupt.edu.cn

Zhifang Le
Shaanxi Normal University, Shaanxi,
710119,China
E-mail: 18016424131@163.com

Shengnan Zhou
Beijing University of Posts and Telecommunications, Beijing,
100876, China
E-mail: 690707520@qq.com

In order to recommend TV programs and advertisements for TV audiences more efficiently, a new method of constructing diversified and accurate TV user profiles is proposed. In our method, the micro-blog users and the TV users will be treated as the same users because they are concerned about the TV programs; hence, the tags of micro-blog users are obtained by the web crawler first, then micro-blog users data are applied to build the model and use this model to predict the TV user's tags. In order to evaluate the accuracy of user profiles, we use thereal viewing logs for a month. Our method is then evaluated by the content-based recommendation system. Experimental results show, compared with other algorithms, our method features better performance in the Precision and Area Under Curve (AUC). Thus, the method of using micro-blog data to construct TV user profiles is an effective solution.

CENet2017
22-23 July, 2017
Shanghai, China

[1]Speaker

PoS(CENet2017)056

1. Introduction

With the rapid increase of the amount of smart TV and the development of communication technologies, the smart TV has become an important device for family and given us an opportunity to recommend the users' preferred TV programs and advertisements in the smart TV. At the same time, different users may have different preferences. It is thus of great importance to construct TV user profiles in order to recommend the content effectively [1-2].

At present, many scholars obtain TV program tags through the Electronic Program Guide (EPG) and use simple statistical method, *tf-idf* and clustering to analyze the relationship between TV programs and TV users to get the TV user profiles [3-4]. For example, assuming that a user has a great probability of watching the news TV program,the user will add tag *news*. However, the above method can only reflect the preference of the user to the TV program rather than the fundamental characteristics of a user (e.g. Age, gender). Many tags from EPG such as *Talk show, Variety show, News, etc.* are too unspecific to reflect the user's key features.

Based on the above problems, we propose a new method of constructing TV user profiles on the basis of the micro-blogM, which has been widely used as a tool for information exchange and sharing, and studied by more and more scholars in recent years [5-6]. Fig. 1 shows that micro-blog users will not only fill in the registration of gender, birthday, location and other basic information, but also fill in the interest preferences tags. Such information filled by the user is the most accurate data to study the user profiles. But so far, no scholar has been using micro-blog for building the TV user profiles.

Figure 1: Micro-blog User's Basic Information and Tag Information

The input data including Micro-blog data and TV viewing logs originate from different users. Although the micro-blog users and TV users are not on the same platform, but this paper treats them as the same because they will concerned more about their favorite TV programs of different forms. For example, a TV user expresses his/her concern about a TV program by watching. Likewise, a micro-blog user expresses his/her concern about a TV program by writing weibo (e.g. #The Journey of Flower#) and the official accounts. Thus, we use micro-blog users' concerning about TV programs as the feature of classifier and use the micro-blog users' tags as the result of classifier, then the predict classification models were constructed by *xgboost*. Because the micro-blog user and the TV user are treated as the same kind of users, we could apply this classification models to the TV users to predict the probability that the user has each tag. As the TV user profiles tags come from the micro-blog user's basic information and tag information in our method, our method is maintaly superior because we have diversified and accurate tags for constructing the TV user profiles.

In order to determine the accuracy of the user profiles, we use a month's real viewing logs produced by TV users. Our method is then tested by the content-based recommendation system. Experimental results show that, compared with other algorithms, our method highlights sound performances in precision and AUC [7].

The rest of the paper is structured as follows. Section 2 discusses related works and the proposed method is introduced in Section 3. Section 4 describes the data set used in this study, the experimental method and the result of the proposed method's evaluation. Finally, Section 5 makes conclusion.

2. Related Work

2.1Research in Micro-Blog

Because of the large number of users and the characteristics of openness, real-time and content diversity, the social network has become a hot topic in recent years. Xu et al. crawled micro-blog data through web crawler, then an algorithm of computing user similarity was designed by analyzing the users' comments, concerns and micro-blogs. The corresponding experiments showed that the users with similar behaviors feature strong correlation [8]. Lin et al. classified the user's emotion by analyzing the comments on micro-blog and the average accuracy rate could reach over 85% in terms of that user's comments on micro-blog which reflects the user's emotional preferences [9]. Recently, some scholars have studied the relationship between micro-blog users and TV programs. Some scholarsstudied the relationship between tweets and TV programs [5]. Wakamiya et at. and Juanjuan et at. accurately predicted the ratings of TV programs by tapping the relationship between micro-blog users and TV programs [6, 10]. According to the above research, we can see that the micro-blog is an important channel for the study of TV user profiles, but no one has yet used micro-blog information to constructing the TV user profiles.

2.2 User Profiles

With the rapid development of the Internet, it is becoming more and more important to construct the users' profiles to understand the their basic information and interests. Boubiche et al. proposed a user profiles model base on search records in order to solve the multi-purpose problem by using the search engine [1]. Furletti et al. and Zhou et at. utilized simple judgment strategy and clustering method to built the user profiles for mobile phone users [2, 11]; however, the simple judgment strategy is highly subjective and the method of clustering in non-supervised learning (e.g. K-means) can't be accurately predicted for some marginal users. Zhao et at. proposed the behavior factorization as a way to built user profiles in micro-blog [12]. The user profiles on micro-blog has achieved good results, but TV user profiles is still in its infancy. Yu et at. and Iguchi et at. collected TV program tags through EPG, then designed strategies from tag merging or audience classification so as to get TV user profiles [3, 13]. Naemura et al. got extra TV program tags by crawling Wikipad [4]. However, these methods can only get the TV programs that the user preferred, but cannot get more basic tags, such as gender and age, etc.

Based on the above problems, this paper explores the relationship between TV users, micro-blog users and TV programs, and uses the classification algorithm to predict the user profiles.

3. Proposed Method

As shown in Fig. 2, our proposed method for constructing TV user profiles comprises two main steps:

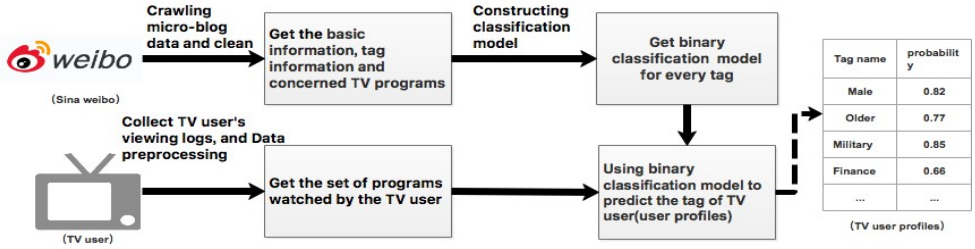

Figure 2: Step of TV User Profiles

- Mining micro-blog data and get binary classification model for every tag.
- TV users' viewing logs preprocessing to obtain the user profiles by binary classifications.

3.1 Data Preprocessing

Given a set of TV programs $P=\{p_1, p_2, \ldots, p_k\}$, and a set of micro-blog users $WU=\{wu_1, wu_2, \ldots, wu_n\}$, the input data from micro-blog can be represented as a set of tuples:

$$\zeta=\{[BaseInfo(wu_i), TagInfo(wu_i), Weibo(wu_i), Followee(wu_i)], i=1,2,\ldots,n\} \quad (3.1)$$

$$BaseInfo(wu_i)=\{Gender(wu_i), Age(wu_i)\} \quad (3.2)$$

where $Gender(wu_i)$ and $Age(wu_i)$ represents the gender and the age of micro-blog user wu_i, $TagInfo(wu_i)$ represents wu_i's Tag information shown in Fig. 1, $Weibo(wu_i)$ represents all weibos posed by wu_i, $Followee(wu_i)$ represents all TV programs in P followed by wu_i. Then a vector ws_i that reflects wu_i's concerned TV programs can be represents as: $ws_i=\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$.

$$x_{ij}=\begin{cases}1; & p_j \in Followee(wu_i) \vee The\ weibo\ contain\ p_j \in Weibo(wu_i) \\ 0; & Otherwise\end{cases} \quad (3.3)$$

we will use ws_i as the feature in training classification model.

Given a set of TV users $TU=\{tu_1, tu_2, \ldots, tu_m\}$, the TV viewing logs can be represented as a $m\times k$ matrix M, M_{ij} represents the time that tu_i watch p_j, then the similarity like ws_i, we can get a vector vs_i that reflects tu_i's concerned TV program: $vs_i=\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$.

$$x_{ij}=\begin{cases}1; & 2\times M_{ij} \geqslant PlayTime(p_j) \\ 0; & Otherwise\end{cases} \quad (3.4)$$

where $PlayTime(p_j)$ represents the total play time of p_j. We will use vs_i as the feature of TV user.

3.2 Build Predictive Models through Micro-blog data

We choose the basic information tags set BT and the interest preference tags set IT as a set of TV user profiles T: $BT=\{Gender, Age\}$, $IT=\{Literature, Sports, Star, Finance, Military, Music, Cate, Film\}$.

$$T=\{BT, IT\} \tag{3.5}$$

In this section, for each tag in T, we will describe how to build a predicted model.

3.2.1 Gender Tag Model Construction

Every micro-blog user must fill in the gender information, therefore, the gender information that the user fills in will be the result of the classification. The binary classification of the gender tag can be expressed as: $S_{sex}=\{(ws_1, ts_1), (ws_2, ts_2), ..., (ws_n, ts_n)\}, ts_i \in \{0,1\}$.

$$ts_i=\begin{cases}0; Gender(wu_i)=Female\\1; Gender(wu_i)=Male\end{cases} \tag{3.6}$$

3.2.2 Age Tag Model Construction

The micro-blog user's age information can be calculated by birthday, but it is not reasonable to choose the age as the classification result, thus we define a threshold ω_{age}. The binary classification of the age tag can be expressed as: $S_{age}=\{(ws_1, ta_1), (ws_2, ta_2), ..., (ws_n, ta_n)\}, ta_i \in \{0,1\}$.

$$ta_i=\begin{cases}0; Age(wu_i)<\omega_{age}\\1; Age(wu_i)\geq\omega_{age}\end{cases} \tag{3.7}$$

3.2.3 Interest Preference Tags Model Construction

The micro-blog users will add the interest preference tags (e.g. sports, music, NBA) to their own accounts when registering, but these tags don't have a uniform standard because they are typed by the user. In order to correspond with IT, we defined $SimilarTagSet(it_i)$ for each tag in IT, for example, if we construct the sports tag model, firstly, we need to define: $SimilarTagSet(Sprots)=\{Basketball, Football, NBA, Kobe\ Bryant, ...\}$.

Then the binary classification of the sports can be expressed as: $S_{sports}=\{(ws_1, tsp_1), (ws_2, tsp_2), ..., (ws_n, tsp_n)\}, tsp_i \in \{0,1\}$.

$$tsp_i=\begin{cases}1; TagInfo(wu_i)\cap SimilarTagSet(Sprots)\neq\varnothing\\0; \qquad\qquad Otherwise\end{cases} \tag{3.8}$$

Similarly, we can construct the remaining interest preference tag model.

Then we use *xgboost* on the basis of *logistic regression* to construct the binary classification model for $S_{sex}, S_{age}, S_{sports}, ...$ [14].

In order to reflect the relevance of micro-blog users's tags and concerned TV programs, we randomly selected 80% micro-blog users as as training set, and the other 20% is used as the test set, then got AUC through *xgboost*. As shown in Fig. 3, that most model's AUC is greater than 0.6, and there is also some models's AUC greater than 0.75 such as the s*ports tag model*, the

m*ilitary tag model* and *the financial tag mode*, indicating there is a strong correlation between the TV programs that micro-blog users attention and the micro-blog users' tags, for example, most of users concerned about NBA and have "*Sports*" tag.

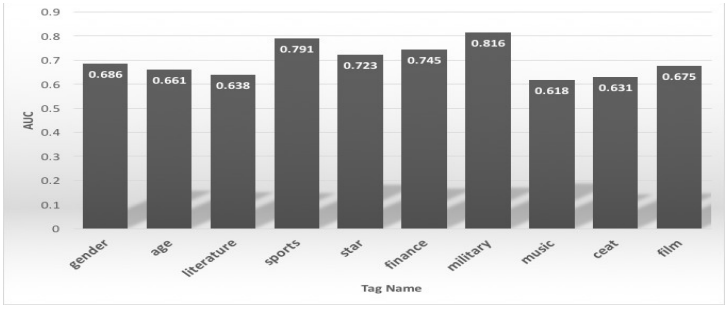

Figure 3: AUC of All Tag Models

3.3 Construct TV User Profiles

As the predicted model was constructed by *logistic regression*, the predicted result is a decimal between 0 and 1. And the feature in TV user vs_i is the same as the feature in the micro-blog user ws_i ,we can use these binary classification model as constructed by Section 3.2 to predict the tag of each TV user. For each TV user, we will get a set of tuples: $TP=\{(t_1,tp_1),(t_2,tp_2),\ldots,(t_{|T|},tp_{|T|})\}$, where t_i is the i -th tag in T , tp_i represents the the probability of containing t_i . As show in Section 3.2.1 and Section 3.2.2, we will get $(Gender, tp(Gender))$ and $(Age, tp(Age))$ for tag gender and tag age, so we need to make changes to the gender tag and the age tag as follows:

$$(Gender, tp(Gender))=\begin{cases}(Male, tp(Gender)); & tp(Gender)\geq 0.5\\ (Female, 1-tp(Gender)); & Otherwise\end{cases} \tag{3.9}$$

where $tp(Gender)=tp_i, t_i=Gender$.

$$(Age, tp(Age))=\begin{cases}(Older, tp(Age)); & tp(Age)\geq 0.5\\ (Younger, 1-tp(Age)); & Otherwise\end{cases} \tag{3.10}$$

where $tp(Age)=tp_i, t_i=Age$, if $tp(Age)$ is more close to 1, we believe that the the users are more possibly older than ω_{age} , so we give the user tag $Older$ and $Younger$ otherwise. In this paper, we use TP as TV user profiles.

4. Experiments

In order to carry out comparative analysis of the user profiles, this paper conducted two evaluation experiments. The first one is based on the EPG data (the method is referred to as Fepg), and the second one used this paper's method (the method is referred to as Fweibo). To test the accuracy of TV user profiles, the experimental method was a recommendation system based on tags, and the assessment methods are precision and AUC. The result shows that the accuracy of Fweibo is markedly higher than Fepg.

4.1 Dataset

The experiment data from:

- About 200,000 TV users' viewing logs for one month.
- Major TV program list during the month.
- 21,000 micro-blog users' information were crawled by the web crawler.
- In order to contrast our method, we got every program's tags in EPG.

4.2 Experimental Method

The tags obtained by the EPG and the tags used in this paper are different. For example, the tags in EPG are: *Talk show, Variety show and News,* etc, but the tags in this paper are: *Age, Gender and Sports,* etc. For the Fepg, the tags are retrieved from the EPG, and then the *tf-idf* is used to label the TV user [15]. For the Fweibo, we will got TP in Section 3.3 for every TV userI. In order to get the exact set of tags CP, we defined a threshold ρ,

$$CP = \{t_i\}, tp_i > \rho, i = 1,2,\ldots,|T| \tag{4.1}$$

That is, all tags with the probability greater than ρ are added to CP, then the tags of the TV program are calculated by counting all the micro-blog user tags that are concerned with the TV program and using the *tf-idf*.

In this paper, we focus on 80% programs that TV users have watched as a feature set and 20% as a validation set. The feature set is used as input, and the Fepg and Fweibo methods are used to obtain the user profiles (a set of tags), then the correlation value between TV users and TV programs is obtained by *cosine similarity*, finally we recommended the TopN TV programs to the TV users.

The precision and AUC are calculated by using the recommended results and the validation set. The accurate rate is calculated by Formula (4.2):

$$precision = \frac{the\ number\ of\ TV\ programs\ that\ recommended\ correctly}{the\ number\ of\ TV\ programs\ that\ recommended} \tag{4.2}$$

We uses an approximate method to obtain AUC [7].

4.3 Results

In order to show the effect of various methods, firstly, we evaluated the method of random recommendations. As a result, we can know the performance of various methods by comparing with the random method. In the experiment of the recommended system, the recommend num TopN is an important factor affecting the accuracy rate, thus we evaluated various methods for TopN=1,2,3,4. As shown in Fig. 4, in comparison with the random method, Fepg features a good performance. Compared to the Fepg, our Fweibo has 142% improvement on average. The combination method Fepg+Fweibo has the highest precision because the TV user profiles from the Fepg are different to the TV user profiles from the Fweibo.

We found that the top 20% TV programs accounting for top 80% of the total viewing time and the distribution of data is very uneven. We use AUC as a standard of evaluation because AUC is not affected by data distribution. As shown in Fig. 5, compared to the Fegp, our Fweibo has 20% improvement and the Fepg+Fweibo has 23% improvement. The improvement between the Fweibo and the Fepg+Fweibo is few because the Fepg has little valuable information in comparison with the Fweibo.

Depending on the above results, the Fweibo that uses micro-blog data to construct TV user profile have a significant performance boost.

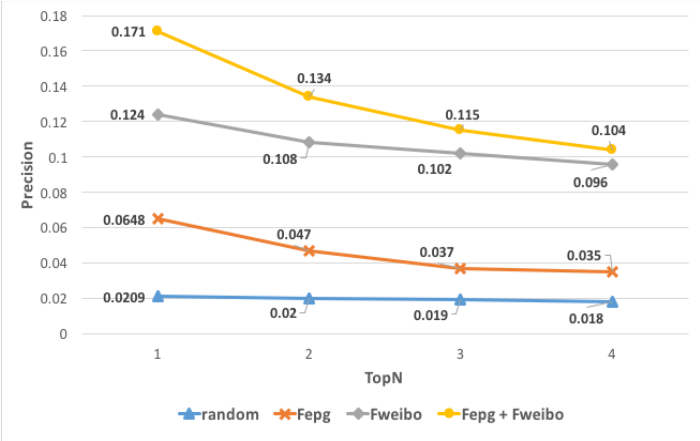

Figure 4: Precision of Various Methods

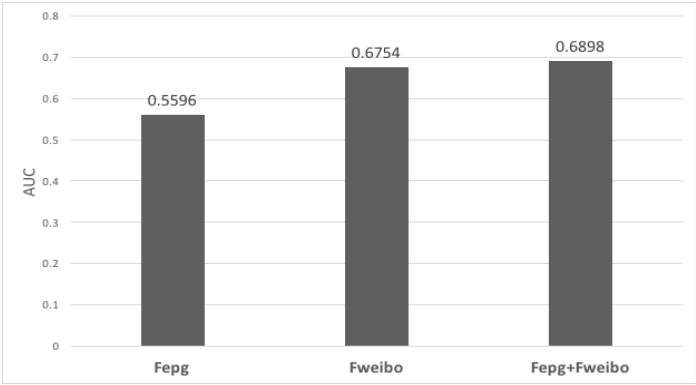

Figure 5: AUC of Various Methods

5. Conclusion and Future Work

This paper proposed a new method for constructing diversified TV user profiles based on micro-blog and designed a recommended experimental method for verification. Firstly, we introduced benefits by using micro-blog data and explained why micro-blog users and television users can be considered as users of the same type. Then we constructed the tag classification model by micro-blog data and used these classification models to get TV user profiles. Finally, we use the data of real viewing logs for one month and the method is then evaluated by the content-based recommendation system. The experimental results show that the method that uses micro-blogging data to construct TV user profile is an effective solution. The following tasks will be achieved in the future:

(1) Population distributions of micro-blog users and TV users are different (e.g. 90% micro-blog users are from twenty to forty years old). Thus, in the future, we will cluster TV users, screen out those whose actions are the same as the micro-blog users' and then generate their user profiles.

(2) User groups of TV users are various because their working time and rest time are different. To improve the accuracy of TV user profiles, the working time and the rest time will be calculated separately.

References

[1] S. Alaoui, Y.E.B.E. Idrissi, R. Ajhoun. *Building rich user profile based on intentional perspective*[C]. International Conference on Advanced Wireless Information and Communication Technologies, AWICT 2015. Procedia Computer Science. Sousse, Tunisia. October. 342-349(2015)

[2] B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo. *Identifying users profiles from mobile calls habits*[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. Beijing, China. August. 17-24(2012).

[3] Z. Yu, X. Zhou, Y. Hao, J. Gu. *Tv program recommendation for multiple viewers based on user profile merging*[J]. User Modeling and User-Adapted Interaction. *16*(1), 63-82(2006).

[4] M. Naemura, M. Takahashi, S. Clippingdale, Y. Yamanouchi, H. Fujisawa. *Constructing personalized user profiles through TV viewing*[C]. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. IEEE. Nara, Japan. June. 7521929(2016).

[5] P. Cremonesi, R. Pagano, S. Pasquali, R. Turrin. *TV program detection in tweets*[C]. 11th European Conference on Interactive TV and Video, EuroITV 2013. ACM, Como, Italy. June. 45-53(2013).

[6] S. Wakamiya, R. Lee, K. Sumiya. *Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV rating*[C]. International Conference on Ubiquitous Information Management and Communication, Icuimc 2011. ACM. Seoul, Republic of Korea. February. 39(2011).

[7] Y.X. Zhu, L.Y. Lü. *Evaluation metrics for recommender systems*[J]. journal of the University of Electronic Science & Technology of China, 41(2), 163-175(2012). (In Chinese)

[8] Z.M. Xu, D. Li, T. Liu, S. Li, G. Wang, S.L. Yuan. *Measuring similarity between microblog users and its application*[J]. Chinese Journal of Computers. *37(1)*, 207- 218(2014). (In Chinese)

[9] Y.M. Lin, T. Zhu, X.L. Wang, A.Y. Zhou. *Assembling and optimizing multiple classifiers for user opinion analysis*[J]. Chinese Journal of Computers. *36*(8), 1650-1658(2013). (In Chinese)

[10] J.J. Tu, T.M. Liu. *A predicting model of tv audience rating based on the decision tree*[J]. Microcomputer Information, 23(27), 251-252(2007). (In Chinese)

[11] Y.Q. Zhou, Y.F Hu, H.C. He. *Learning User Profile in the Personalization News Service*[C]. International Conference on Natural Language Processing and Knowledge Engineering. IEEE Xplore. Beijing, China. August. 485-490(2007).

[12] Z. Zhao, Z. Cheng, L. Hong, E.H. Chi. *Improving user topic interest profiles by behavior factorization*[C]. 24th International Conference on World Wide Web, WWW 2015. ACM. Florence, Italy. May. 1406-1416(2015).

[13] K. Iguchi, Y. Hijikata, S. Nishida. *Individualizing user profile from viewing logs of several people for TV program recommendation*[C]. 9th International Conference on Ubiquitous Information Management and Communication, ACM IMCOM 2015. ACM. Bali, Indonesia. January. A61(2015).

[14] T.Q Chen, C. Guestrin. XGBoost: *A Scalable Tree Boosting System*[C]. KDD'16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. New York, USA. August. 785-794(2016).

[15] C.H. Huang, J. Yin, F. Hou. *A text similarity measurement combining word semantic information with tf-idf method*[J]. Chinese Journal of Computers,34(5), 856-864(2011). (In Chinese)