User Modeling on Social Networks——Using User Tags and Weibo Content for User Modeling

Xuehua Chi[*]

Nanjing University of Science & Technology; Nanjing; Jiangsu; China
[*]Corresponding author

Abstract—**User-generated content is of great significance for user modeling and user interest mining. This paper defines a microblog user model which combining weibo content and tags with vector space model(VSM) representation. The user model consists of two parts, one part, user interest representation based on weibo content: pretreatment, feature extraction, and then compute the characteristic value with TF-IDF method, after that, the user's weibo content is expressed by VSM; Another part, user interest representation based on user tags: feature extraction and word frequency method for computing the characteristic value and user tags are expressed by VSM. Finally, the resulting user model can be obtained by combining the two parts mentioned above.**

Keywords—user modeling; user tag; vector space model; feature extraction

I. Introduction

The development of the Internet leads to the massive growth of data and make people spend more time on search of information. But users prefer to enjoy the efficient, personalized service such as recommending topics of interest to users and more accurate advertisement content. In personalized information services, the research on user modeling technology has become the key technology of personalized service.

Today, social networking has become popular, and social platforms like Weibo, WeChat, Douban are very common, among which Weibo is China's largest social media and plays an important role in people's daily life. There are large quantity of available user-generated contents (UGCs) on the Weibo which contain huge information with high quality. What users have browsed and issued on this social platform reflects their concerns. And now more and more social platforms provide label function, which users can not only label for the resource, but also label for themselves. These user tags which reflect user's attributes and interests are also of great significance for user modeling and user interest mining. In this paper, a method of constructing user model is introduced by combing user's weibo content and tags.

II. Related Work

A rich set of studies has been conducted on user modeling. By information acquisition in user modeling, the modeling can be divided into display modeling and implicit modeling [1]. The quality of the user model is closely related to the model representation. At present, common user model representations are: keyword list method representation [2], topic representation [3], neural network-based representation [4], ontology-based representation [5] and representation based on vector space model [6]. In addition, different types of data like text, picture, social networks, etc. are used for user modeling [7-9].

The methods of constructing user's interest model on the social network are also proposed by some researches, such as extracting microblog keywords on the Weibo platform to find user interests [10][11], using the user's interaction relationship for user tag recommendation[12][13] , mining the interest of the user by topic model [14].

Most of the user modeling focuses on information collection. But it is difficult to ensure the integrity of user information. We can use information that have strong representation of user characteristics to improve the efficiency and quality of user modeling. On Sina Weibo, the user tag and weibo content reflect user's preferences. Therefore, this paper will mine the user's interest from the perspective of combining the user tag and the Weibo content, and build a user model.

III. Using Tags and Weibo Content for User Modeling

A. Method for User Model

In our approach as seen in Figure 1., user model consists of two parts, one part, user interest representation about weibo content; another part, user interest representation about the user tags.
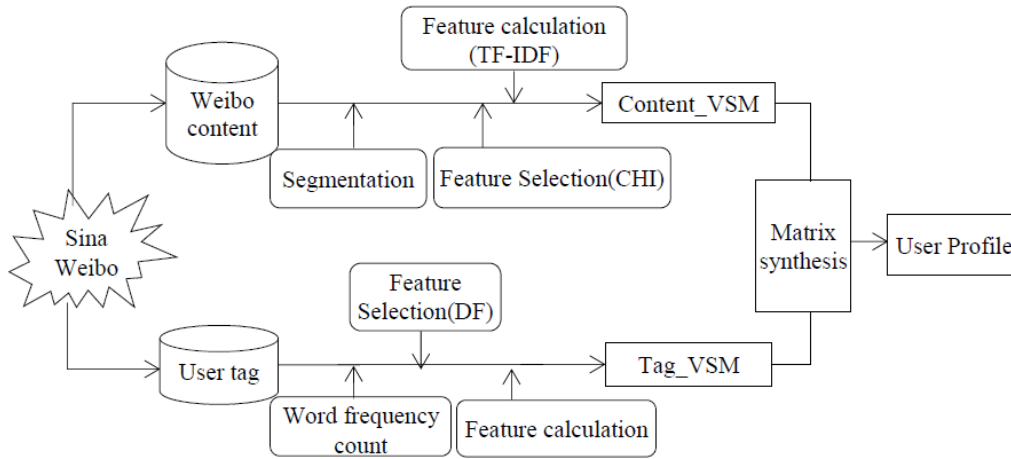
FIGURE I. USING TAGS AND WEIBO CONTENT FOR USER MODELING

Firstly, user interest representation about weibo content: pretreatment like Chinese word segmentation, feature extraction, and then compute the characteristic value with TF-IDF method, after that, the user's weibo content is represented by vector space model (VSM);

Secondly, user interest representation about the user tags: document frequency statistics method for feature extraction and then computing the characteristic value with the word frequency count method and the user tag is also represented by VSM.

Thirdly, the resulting user model can be obtained by combining the two parts mentioned above.

B. User Model Representation

The representation of the user model determines the computability of user model, and it also limits the selection of user modeling methods. Vector Space Model (VSM) is a statistical model that is widely used in information retrieval and it is effective to represent text topics [15]. This paper uses the vector space model to represent user weibo content and user tags.

IV. EXPERIMENT

A. Dataset and Evaluation

1) Dataset: For this study, we collected 4689 verified users from Sina Weibo. As a summary, there are 814,516 user posts and 20,996 tags with different domains in Table 1.

TABLE I. DATASET DISTRIBUTION OF DIFFERENT DOMAINS

Domain	User Number	Microblogs Number	Tags Number
internet	743	159218	5691
law	556	70764	2412
literature	1693	331243	8039
medical	805	111193	2314
football	892	142098	2540
Total	4689	814516	20996

2) Evaluation.: Each user is modeled and we invite users to feedback the score of the model. The value of satisfaction is calculated as follows:

$$S = \frac{SUM(G_i)}{N} \times 100\% \qquad (1)$$

S represents degree of satisfaction, G_i represents model score of user i, N represents user number. The degree of satisfaction is show in table 2.

TABLE II. THE DEGREE OF SATISFACTION

Score	10~9	8~7	6~5	4~3	2~0
degree	very satisfied	satisfied	general	not satisfied	very dissatisfied

B. Results Analysis

The final model consists of TOP-20 content feature words and user tags. Table 3. details five users' model randomly selected from all users.

TABLE III. THE SAMPLE OF USER MODEL

User ID	Content_VSM_top20	Tag_VSM
1××××××××3	Jianwang, Xi shan ju, engine, programmer, 3D, code, API, multithreading, Pphase 3, FPS, player, training, optimized, DirectX, memory, gossip, version	badminton, humor, photography, online, programmer
1××××××××0	netease, singles, wind, movie tickets, music, E-mail, yours,news, club, mraz, lyrics, Guangzhou, include, Yixin, discount shop, Mo zi, free	investment and financing, scorpio man, shopping
1××××××××7	E-commerce, brand, Jingdong, e-commerce, marketing, consumers, distributors, profit-making, department, sleep, cost team, operation	Online shopping, brand, public relations, creativity, IT
2××××××××3	lawyer, audit, Li Changkui, Xiangtan, Chen Ruihua, Chen Weidong, judge, criminal law, business tax, public trial, presumptive, practice, defense, value-added tax, principal, weaving, criminal procedure law	Criminal, Lawyer
1××××××××4	song and dance duet ,tag, lawyer, discrimination, ATM, perjury, Beihai, Guangxi, SARFT, bank, insulation materials, Beijing, criminal detention, risk awareness, counsel	law, Beijing

User tags can be divided into the following categories: user's own attributes, including constellation, blood type, personality, etc.; user's interests and hobbies; user professional. Most of these tags are too broad, especially those about professional, such as tag "law". But "law" is just a generic name which includes

Criminal Law, Civil Procedure Law, Administrative Law, Trademark Law, etc. Therefore, we do not know which law user belongs to.

Fortunately, Weibo content reflects the user's recent status and concerns, and is a supplement to user tags. For example, user "1××××××××4" detailed more such as "crime detention" and "forgery" in weibo content, who's tag is "Law". Therefore, the combination of the user's Weibo content and the user's tag will show the user profile clearly, which help to make more accurate recommendations.

In the model evaluation, we randomly selected several users from five domains. Figure 2 shows that 465 users have been invited for feedback the model score and 36 have responded effectively. After statistics, the satisfaction results are shown in Table 4.
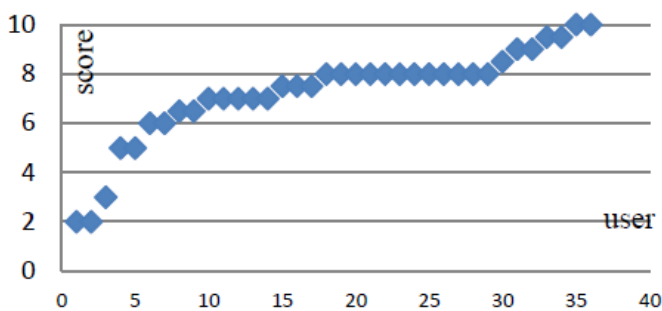

FIGURE II. SCORE OF MODELS

TABLE IV. ATIFACTION VALUE OF USER MODEL

Domain	Satisfaction Value	Number of Invited Users
internet	7.14	100
law	8.3	100
literature	7.1	100
medical	6.75	85
football	8	80
Total	7.25	465

The total satisfaction value is 7.25 points, which indicating the user modeling is effective. The domain of football has the lowest value of satisfaction, followed by the Internet and literature. Medical and Law have higher value. The possible reason for the low satisfaction of the football user model is that users microblogs not only care about sports events but also involve other contents. These users' contents may be more complicated, leading to less prominent of interest. Medical and law users have high degree of satisfaction because of high related professional by contrast.

V. CONCLUSION

Sina Weibo is China's largest social media and plays an important role in people's daily life. This paper collects data from Sina Weibo for user modeling. User model consists of two parts: user interest representation based on weibo content and the user tags in Space Vector Model, and the TOP-20 feature vectors and user tag vectors are selected to represent the user's final model. This representation clearly expresses user interests, and the standardized vectors also facilitate the completion of tasks in subsequent phases.

The user's label expresses the user in a short and powerful manner, including its own attributes, professions, and hobbies. Weibo content reflects the user's concerns, which is a refinement and supplement of user tags. The result of evaluation indicating that the user model combining user content and tags is of high quality.

However, in this paper, it is not sufficient to combine user microblogs and tags in semantic like words can be integrated with the same conceptual. In addition, the evaluation methods for user models have yet to be improved.

REFERENCES

[1] M. J. Yi, Research on User Modeling for User Modeling for Personalized Services, Journal of Intelligence, vol.3, pp. 77-79, 2006.

[2] Y. L. Zhang, and Wang Quan. User Profile Mining of Combining Web Behavior and Content Analysis, New Technology of Library and Information Service, vol.6, pp. 52-55, 2007.

[3] A. Ahmed, Y. Low, M. Aly, et al. Scalable distributed inference of dynamic user interests for behavioral targeting, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 114-122, 2011.

[4] C. Lu, W. Lam, and Y. Zhang, Twitter user modeling and tweets recommendation based on wikipedia concept graph, Proceedings of Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 33-38, 2012.

[5] G. Semeraro, M. Degemmis, P. Lops, et al, Combining Learning and Word Sense Disambiguation for Intelligent User Profiling, Proceedings of the International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 2856-2861,2007.

[6] W. T. Chen, X. M. Zhang, and Z. J. Li. Analysis of Topic Models on Modeling MicroBlog User Interestingness, Computer Science, vol.4, pp. 127-130, 2013.

[7] A. Tiroshi, Graph based user modeling, Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, New York, NY, USA, pp.371-374, 2012.

[8] D. Brickley, and L. Miller, FOAF Vocabulary Specification 0.91[EB/OL].[2012-03-01].

[9] H. N. Kim, A. Alkhaldi, A. ElSaddik, et al, Collaborative user modeling with user-generated tags for social recommender systems, Expert Systems with Applications, vol.7, pp.8488-8496, 2011.

[10] W. Wu, B. Zhang, and M. Ostendorf, Automatic generation of personalized annotation tags for twitter users, Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.689-692, 2010.

[11] Y. Yamaguchi, T. Amagasa, and H. Kitagawa, Tag-based User Topic Discovery Using Twitter Lists, Proceedings of the 2011 international conference on Advances in Social Networks Analysis and Mining, pp.13-20, 2011.

[12] J. L. Zhang. Application of Label Propagation Algorithm in Weibo User Interest, Programmer, vol.7, pp.50-53,2012.

[13] J. Wen, P. E. Lim, J. Jiang, et al., Twitter Rank: finding topic-sensitive influential twitters, Proceedings of the third ACM international conference on Web search and data mining, pp.261-270, 2010.

[14] B. Yatus, and R. NetoB, Modern Information Retrieval, China Machine Press, 2011.

[15] D. Munteanu, Vector Space Model for Document Representation in Information Retrieval, 2007.