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.
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>
<thead>
<tr>
<th>Domain</th>
<th>User Number</th>
<th>Microblogs Number</th>
<th>Tags Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>internet</td>
<td>743</td>
<td>159218</td>
<td>5691</td>
</tr>
<tr>
<td>law</td>
<td>556</td>
<td>70764</td>
<td>2412</td>
</tr>
<tr>
<td>literature</td>
<td>1693</td>
<td>331243</td>
<td>8039</td>
</tr>
<tr>
<td>medical</td>
<td>805</td>
<td>111193</td>
<td>2314</td>
</tr>
<tr>
<td>football</td>
<td>892</td>
<td>142098</td>
<td>2540</td>
</tr>
<tr>
<td>Total</td>
<td>4689</td>
<td>814516</td>
<td>20996</td>
</tr>
</tbody>
</table>

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:

\[ S = \frac{\sum (G_i)}{N} \times 100\% \]  

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

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.

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 medical with the same conceptual. In addition, the evaluation methods for user models have yet to be improved.

**ACKNOWLEDGMENT**

This work is supported by Postgraduate Research & Practice Innovation Program of JiangSu Province(SJCX17_0094).

**REFERENCES**


