A Novel Interest Mining Method based on User Relationship-topic Model

Peng Wang, Cheng Gao*, Huamin Yang, Songjiang Li, Ningjia Qiu

School of Computer Science and Technology, Changchun University of Science and Technology
ChangChun, 130002, China, 0431-85583583
gaocheng_dove@126.com*

Abstract — The modelling of user interest is an important process of web personalized recommendation. To make recommendations to the users based on the collaborative filtering of web browsing history only meets their current needs, it does not satisfy the users' long-term interests and ignores the effect of interests in relationships in the social network. Aiming at this problem, this paper proposes a method of interest mining based on user relation-topic model. According to the model of chosen users' topic, we calculate the corresponding interest distribution, and compute user's influence in social networks by the choice relationship. Finally, the target user's interest distribution is weighted. The experimental results show that the average coverage and the prediction accuracy have accurate values, and that the model is effective.

Keywords - interest mining; relationship-topic model; user influence; personalized recommendation

I. INTRODUCTION

With the rapid development of computer network, the increasing popularity of the mobile Internet, the explosive development of various social networking platforms, weibo, as a kind of information sharing and communication platform, is widely loved by the people. As of December 2014, the number of Chinese weibo users is 249 million and the mobile phone weibo users for 171 million [1]. Through the microblog, people share and focus on things that they are interested when and where, which has become the norm. In the era of big data, according to the user's behavior data, to excavate the potential users' interests for understanding the social structure distribution, social network marketing and the personalized recommendation has the very vital significance.

On the study of digging users' interests, AkshayJava[2] analyzed the characteristics of the data source of the Twitter users’ interests, put forward the methods on analyzing the users’ interests according to the variety of datum. This method has certain limitation because of the complexity of the user’s data. HaewoonKwak[3] studied Twitter users’ network topology, based the users’ participation degree on the hot topics to reflect user's interest. Matthew Michelson and Sofus A. Macskassy[4] presented to discuss the characteristic of the user's interest in terms of the text data issued by the user to build the topic model, which did not take into account the influence of the external followed users, so there exist certain defects. Deng Ailin[5] put forward to estimate the target user's interest according to user's predicts to the projects’ scores and the similarity between the users and their neighbors, and then made the recommendation. His method solved the data extreme thin caused by the users’ datum loss, at the same time, it existed the problem that the model effect has a bigger differences along with the change of the neighbor users. Hu Wei[6] proposes a model of interest based on social tagging, which reflected the user's interest direction through the interaction of the users and grading project labels. The accuracy of this method depends on the accuracy of the project labels. On the basis of predecessors' research, aimed at the problems of mining weibo users' interests, this paper proposes a method on interests mining based on the users’ relationship subject model.

By using the followed relationship on the weibo users, as for the text data which the followed users released, this method applies to LDA (Latent Dirichlet Allocation[7-9], implicit Dirichlet distribution) model to calculate their interest subject distribution, then based on the followed relationship, makes use of influence model to compute the weights of the users and further weights the interest topics of the target users. Model process as shown in the following figure:
II. INTEREST TOPIC MODEL

To mine the users’ interests, it begins to determine the data set which could reflect the users’ interests. The users produce a large amount of behavior datum by many forms such as publishing information, reviews, following friends, browsing the web and so on. In these many behavior datum, the text data and the following datum of the users are the expression of the users’ subjective intention. They both can most directly response to the characteristics of their interests. Therefore, this paper presents a comprehensive model based on relationship - theme, by focusing on the user’s text data and the follow relationship of target users to predict the interest topic of the target users.

LDA is a bayesian model of three layers, which is made of the document layer, topic layer and the key layer, each layer of which has the corresponding random variables or parameters to control[10]. Its basic idea is that the text is generated by mixing crytic themes at random, each topic are distributed for corresponding a particular key. LDA model assumes that all the documentation have K implied theme, to generate a document, it firstly generates a subject distribution of the document and then generates a collection of words; To generate a word, it needs to randomly choose a topic according to the theme distribution of the document, and then randomly selects a word according to the words distribution of the theme, and further repeats the process until the document is generated.

![LDA Probabilistic Graphical Model](image)

Fig.2 is this schematic process of the LDA model, of which K is the theme, M is the total number of documents, Nm is the total number of words in the mth document; β is the Dirichlet prior parameters of the keys’ multinomial distribution under each topic, α is the Dirichlet prior parameters of the themes’ multinomial distribution under each document. zm,n is the theme of the nth word in the mth document, wth is the nth word in m document. The both remaining implicit variables, namely θm and φk, respectively means the subject distribution in the mth document and the keys’ distribution in the kth theme. The former is k dimensional vector (k for the total number of the topic), and the latter is the v dimensional vector (v for the total number of words in the dictionary).

Given a collection of documents, wth is the known variable which can be observed. α and β are the prior parameters given by the experience. The other variables, namely zm,n, θm and φk are the unknown hidden variables, which also need to be learnt and estimated according to the observed variables. According to the figure of the LDA model, we can write the joint distribution of all variables. The joint probability of themes and key words are as follows:

$$P(\theta, z, w | \alpha, \beta) = P(\theta | \alpha) \prod_{n=1}^{N} P(z_n | \theta) P(w_n | z_n, \beta)$$

Integrate θ and z to get the edge distribution of the document:

$$P(\theta | \alpha) = \int P(\theta | \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} P(z_n | \theta) P(w_n | z_n, \beta) \right) d\theta$$

In the LDA model, the estimated text of parameters uses the Gibbs Sampling algorithm. The Gibbs Sampling is a special case of MCMC[11] (Markov Chain Monte Carlo) algorithm. The idea of this algorithm is that it selects a dimension of the probability vector for each time, then gives the variable values sampling of other dimensions to determine the value of the current dimension, does the iteration until convergence output the estimated parameters. The specific process as shown in the figure below:
In the beginning, we randomly distribute the themes for each word in the text, and then count the number of keys appeared in each topic and the number of keys of the z theme in each document, the calculation of each round round \( P(z_i \mid z_{-i}, d, w) \), means excludes the theme distribution of the current word, distributes and estimates the probability of each topic distributed by the current words according to the theme of all the other words. When getting that the current word belongs to the probability distribution of all subject \( z \), we randomly select a new subject \( z(1) \) for this word according to this probability distribution, then constantly update the theme of the next word with the same method, until find the convergence of the theme distribution \( \theta_m \) in each document and the key distribution \( \phi_k \) in each theme, then stop the algorithm, and output the being estimated parameters \( \theta_m \) and \( \phi_k \), finally get the theme \( \theta_m \) of each word at the same time. In the actual application, we will set the maximum number of iterations. The calculation formula for each time is called the Gibbs Update Rules. The computation formula is as follows:

\[
\begin{align*}
P(z_i = k \mid z_{-i}, w) &= \frac{P(w, z_i = k)}{P(w, z_{-i})} \\
&= \frac{P(w \mid z_i = k) \cdot P(z_i = k \mid z_{-i})}{P(w \mid z_{-i})} \\
&= \frac{n_{i,k}^{(t)} + \beta_{k}}{\sum_{k=1}^{K} n_{i,k}^{(t)} + \beta_{k}}
\end{align*}
\]

(3)

When having convergence from Gibbs sampling, we concentrate the theme distribution of all the words to calculate \( \theta_m \) and \( \phi_k \) according to the final document, as the implied variables of the having been estimated probability graph model. The posterior distribution of themes in each document and the posterior distribution of the keys in each topic are as follows:

\[
\begin{align*}
P(\theta_m \mid z_m, \alpha) &= \frac{1}{Z_{\theta_m}} \prod_{n=1}^{N} P(z_{m,n} \mid \theta_m) \cdot P(\theta_m \mid \alpha) \\
&= \text{Dir}(\theta_m \mid n_m + \alpha) \\
P(\phi_k \mid z, w, \beta) &= \frac{1}{Z_{\phi_k}} \prod_{i \in z} P(w_i \mid \phi_k) \cdot P(\phi_k \mid \beta) \\
&= \text{Dir}(\phi_k \mid n_k + \beta)
\end{align*}
\]

(4)

(5)

From it, we can see that the two posterior distributions and the corresponding prior distribution are the same, are still the Dirichlet distribution, which is decided by the nature of the conjugate distribution. By using the expectations formula of the Dirichlet distribution, we can get the estimated values of two polynomial distribution parameters \( \theta_m \) and \( \phi_k \):

\[
\hat{\phi}_{k,i} = \frac{n_{i}^{(t)} + \beta_{i}}{\sum_{i,t} n_{i}^{(t)} + \beta_{i}} \quad \hat{\theta}_{m,k} = \frac{n_{m}^{(t)} + \alpha_{k}}{\sum_{k=1}^{K} n_{m}^{(t)} + \alpha_{k}}
\]

(6)

III. THE INDEX OF THE USERS’ INFLUENCE

Weibo is a platform for information dissemination and sharing. The influence of the users directly shows the ability to information transmission and sharing. Weibo’s communication ability is directly related to the degree of users’ being followed, at the same time, the influence of the users also has a lot to do with the interest in the topic, the users usually have a big impact in the interest of hobby. So this article puts forward two kinds of the users’ influence index: PeopleRank index, Topic-PeopleRank indicator, the two of which can better respond to the influence ability among users and the users’ influence ability under different themes, the indicators are more detailed.

A. PeopleRank index

PeopleRank indicator is constructed by the principle of PageRank[12]. Users’ follow are seen as the outlink, fans (be followed) as the inlink. Assuming that the greater the number of users’ fans, the more important the users; the more important the user, the more important his followed users, that is, a user’s score is determined by the importance of his fans. The computation formula is as follows:

\[
PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(i)} \frac{PR(p_j)}{L(j)}
\]

(7)

As for the formula, \( PR(p_i) \) stands for the PeopleRank value of the user \( p_i \), \( N \) is the total number of users, \( M(i) \) means the collection of the follow users, \( L(j) \) stands for the number of the users who the user \( p_j \) followed; \( d \) is the damping coefficient, taking value interval \( [0, 1] \), generally the experience value 0.85.

According to the hypothesis, the calculating process of PeopleRank is divided into two steps: firstly, set initially the same PeopleRank value for users in the relationship network and take the PeopleRank value of the users averagely assigned to the out link of the users (i.e., the follow users); Secondly, sum the inlinkPeopleRank value of each user, then users get the ultimate PeopleRank value. In this paper, the user’s initialization PeopleRank value is set to 1, the calculation formula can be simplified as:

\[
PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(i)} \frac{n_{\text{fans}}(j)}{n_{\text{follow}}(j)}
\]

(8)

\( n_{\text{fans}}(j) \) and \( n_{\text{follow}}(j) \) are respectively stands for the fans and follows of the user \( p_i \).

B. Topic-PeopleRank Index

Topic-Sensitive PageRank Haveliwala [13], put forward by Haveliwala in 2002, is the website sorting method aiming
at the content and theme. The method improves the relevance between the PageRank and the Topic. Its basic idea is to calculate the PageRank score of the web pages for different topics. Based on this idea, this paper proposed a Topic-PeopleRank index. Assuming that the probability vector of the user’s interest topic is \( t \), then the Topic-PeopleRank expresses the PageRank value of the user on each topic \( t \). The calculation steps are divided into three steps: firstly, take advantage of the LDA model to get the user's interest topic distribution; secondly, calculate each user's PageRank values; thirdly, synthesize the subject probability and PageRank value. The computation formula is as follows:

\[
TPR(p_j) = PR(p_j) = (1-d) \frac{S_i}{|S_i|} + d \sum_{p_j \in M(i)} \frac{PR(p_j)}{L(j)}
\]  

(9)

For it, \( TPR(p_j) \) represents the Topic-PeopleRank value of the user \( p_j \), \( S \) stands for the symbol vector of the user on a certain topic, and \( S_i \) is the symbol \{0,1\} of the user \( p_j \) on a certain topic.

IV. USERS’ INTEREST PREDICTION AND EVALUATION

Through making use of the theme analysis on the follow user’s text data, we get the corresponding subject probability distribution \( P \). Target users have different liking for the interests, which directly reflect in the influence of the follow users. The more interested in a subject, the more authority his follow users in this aspect. Therefore, it has the different weight of the influence between the follow users and the target users’ interests. In this paper, we take the user's influence index as the target user's interest to weighted average, then to get the target user's interest distribution.

According to the index of the users’ influence in this paper, we calculate to get the influence distribution \( F \) of the follow users (i.e.PR or TPR distribution in the above), make the two indicators through the inner product calculation to get distribution \( G \) of the interest subject which the follow users influence on the target user. The formula is as follows:

\[
G(p_j) = P(p_j) \times F(p_j)
\]  

(10)

Through the weighted average of the follow users, we get the interest topic distribution of the target users. The calculation formula is as follows:

\[
P(p_j) = \sum_{p_j \in M(i)} G(p_j)
\]  

(11)

In the evaluation of the predicted results, this paper is to judge the effect of the model by the degree of coverage \( C \) between the user’s personal tags and the predicted results. The degree of coverage \( C \) is the proportion of the user's predicted label to the actual label. Intuitively, the higher the degree of coverage is, the better the results are, and the range is between \([0, 1]\). Using the terms of the theme to determine the topic direction, for the given target user's interest subject distribution \( P \), this paper sets the threshold \( \delta \), when the subject probability \( P_i < \delta \), we will eliminate the interest theme, that is, the probability of this theme is zero. Taking into account the individual label may be lack of some interest characteristics as well as the diversity of prediction, we propose to use the cosine similarity \( Sim \), which comes from the symbol vector of the distribution of the tags and topics, to measure the forecast accuracy. The computation formula of two kinds of evaluation index is as follows:

\[
y_i = \begin{cases} 
1 & P_i \geq \delta \\
0 & P_i < \delta 
\end{cases}
\]  

(12)

\[
C = \sum_{y_i \in label} \frac{n_{label}}{|Y|}
\]  

(13)

\[
Sim(Y, Y_{label}) = \frac{Y \cdot Y_{label}}{|Y||Y_{label}|}
\]  

(14)

V. EXPERIMENTAL ANALYSIS

The Experimental data crawls Weibo users in a community by web crawlers. The Data contains 231 users and their follow users. We crawl the users’ terms including Weibo name, follow number, fans number and users’ labels. We use R software to preprocess the experimental date and delete the data which the tag is empty; finally, our experimental data contains 3982 users. The experimental procedure is divided into four steps: firstly, calculate the user's theme probability distribution; secondly, calculate the user's PeopleRank value and Topic-PeopleRank value; thirdly, calculate the distribution of the target user's interest topic; fourthly, evaluate the predicted results. First of all, we make the word segmentation for the users’ text labels and eliminate the Chinese conventional pause word. After preprocessing, each review corresponds to a word vector, and then we use the TF-IDF algorithm for the corpus set to make the dimensionality reduction, finally deal with the corpus set taking advantaging of the LDA model. We select the prior hyper-parameters \( \alpha \) and \( \beta \) of the LDA model as the experience value \( \alpha=4000/K, \beta=0.01[14] \). In this paper, we test the different subjects’ number \( K \), choose the highest accurate K value, this paper chooses \( K=5 \). The probability distribution of user interest subject in the LDA model is shown in Table 1.

<table>
<thead>
<tr>
<th>Topic-Terms</th>
<th>follow1</th>
<th>follow2</th>
<th>follow3</th>
<th>follow4</th>
<th>follow5</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital.system.food.software.excel</td>
<td>0.1887</td>
<td>0.1818</td>
<td>0.1864</td>
<td>0.2203</td>
<td>0.2203</td>
</tr>
</tbody>
</table>

DOI 10.5013/IJSSST.a.17.33.16
ISSN: 1473-804x online, 1473-8031 print
From table 1, the Topic-Terms are listed according to the key combination of the user's interest theme and the representation theme. The order of the keys is shown from small to big according to the probability of this term representing the subject. The follow column shows the probability distribution of the follow users in different themes. According to the user's follow relationship and the probability distribution of the interest topic, we calculate PeopleRank value and the Topic-PeopleRank value. In order to facilitate comparison and calculation, this article will do the normalization processing for the data, some datum are shown in table 2.

### TABLE II. THE INDEX OF SOME USERS’ INFLUENCE

<table>
<thead>
<tr>
<th>Follow</th>
<th>T:PR</th>
<th>T:PR</th>
<th>T:PR</th>
<th>T:PR</th>
<th>PR:PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow1</td>
<td>0.01468</td>
<td>0.02738</td>
<td>0.00146</td>
<td>0.02688</td>
<td>0.01468</td>
</tr>
<tr>
<td>follow2</td>
<td>0.02418</td>
<td>0.03688</td>
<td>0.02418</td>
<td>0.03848</td>
<td>0.02418</td>
</tr>
<tr>
<td>follow3</td>
<td>0.02239</td>
<td>0.03509</td>
<td>0.03469</td>
<td>0.02239</td>
<td>0.03669</td>
</tr>
<tr>
<td>follow4</td>
<td>0.09359</td>
<td>0.09449</td>
<td>0.09409</td>
<td>0.08179</td>
<td>0.08179</td>
</tr>
<tr>
<td>follow5</td>
<td>0.01377</td>
<td>0.01467</td>
<td>0.00197</td>
<td>0.00197</td>
<td>0.00197</td>
</tr>
</tbody>
</table>

Through the interest topic distribution and influence of the fans, we predict probability distribution of the users’ interest, and make normalization processing to the results. The predicted results of some users are shown in table 3.

### TABLE III. THE PROBABILITY DISTRIBUTION OF THE PREDICTED USERS’ INTEREST SUBJECT

<table>
<thead>
<tr>
<th>Topic-Terms</th>
<th>follow1</th>
<th>follow1</th>
<th>follow1</th>
<th>follow1</th>
<th>follow1</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital.system.food.software.excel</td>
<td>0.226</td>
<td>0.158</td>
<td>0.197</td>
<td>0.185</td>
<td>0.183</td>
</tr>
<tr>
<td>learning.machines.computing.recommendation.sas</td>
<td>0.189</td>
<td>0.164</td>
<td>0.207</td>
<td>0.174</td>
<td>0.234</td>
</tr>
<tr>
<td>data analysis.statistics.travel.life.music</td>
<td>0.207</td>
<td>0.228</td>
<td>0.189</td>
<td>0.214</td>
<td>0.170</td>
</tr>
<tr>
<td>internet.data.analysis.businessintelligence.read</td>
<td>0.189</td>
<td>0.207</td>
<td>0.188</td>
<td>0.243</td>
<td>0.186</td>
</tr>
<tr>
<td>data mining.venture.natural.language._lm.advertisement</td>
<td>0.189</td>
<td>0.243</td>
<td>0.219</td>
<td>0.184</td>
<td>0.227</td>
</tr>
</tbody>
</table>

According to the probability of the predicted users’ interest subject, this paper, by setting the threshold value \( \delta \), make the theme probability vector abstraction as the symbolic vector, and then calculate the average coverage \( C \) of the users and the prediction accuracy \( \text{Sim} \), the results as shown in figure 4.
The selection of threshold value $\delta$ is related to the theme probability of the users and the lower threshold value will lead to too much interest theme. Although there appear the higher coverage $C$ to a certain extent, its accuracy is greatly reduced; if the threshold is too high, it will lead to lack of interest, and also affect the precision. Seen from the table 1, when $\delta=0.1$, there is the higher tag coverage $C=0.81$, but its accuracy only has $\text{Sim}=0.65$. Referring to the coverage and accuracy under different threshold, this paper chooses the threshold value $\delta=0.2$, which the average coverage and precision values are in a higher level, so the model is reasonable and the effect is better.

VI. CONCLUSION

This paper presents a model of predicting users' interests based on the user's social network relationships and text topic analysis. We calculate the interest of users and the influence to predict the user's interest topic, and take advantage of the user's label coverage and prediction accuracy to measure the good or bad of the model. The Experiments show that the proposed model is reasonable and effective. A user label is used to analyze the interest subject of users, because the number of the texts is relatively small and most of them are short text words, so the accuracy of the user's interest subject can be further improved.

REFERENCES