International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.7 (2015), pp.157-170
http://dx.doi.org/10.14257/ijmue.2015.10.7.17

Entity Relationship Modeling Approach Based on Micro- Blog Tag

Junjiang Li[1,2], Chunlin Li[1,2], Youlong Luo[3],Yahui Zhao[2] and Xijun Mao[2]

[1]*State Key Laboratory of Digital Publishing Technology, Beijing, P.R. China*
[2]*School of Computer Science, Wuhan University of Technology, Wuhan, P.R. China*
[3]*Management School, Wuhan University of Technology, Wuhan, P.R. China*
245786447@qq.com, chunlin74@aliyun.com, 1069685498@qq.com, 1140614849@qq.com, 404027587@qq.com

Abstract

Due to the huge information, short length and noise data, the traditional method has poor effect on micro-blog entity relationship modeling. In this paper, a new micro-blog user interests discovering approach based on tag is presented to improve the entity relationship modeling. First the matrix of user tag built by traditional way may generate the problem of sparse matrix in tag recommendation, so we introduce the information of micro-blog and establish the bipartite graph of User-Tag and Tag-Word respectively, then use them to recommend tag to micro-blog users. Meanwhile interactive relationship between users also show their interests, we establish a graph of tag relation by users' relationship and propose a method called Tag Rank on the basis of this graph to improve the precision of the model. Finally, we combine the two methods to discover user interests. In the experiment, we use several measurement metrics: F-value, precision and the recall rate. It is proven that the new approach in the paper have a perfect performance.

Keywords: *Micro-blog; User interest model; Tag; Interest discovering; bipartite graph*

1. Introduction

With the advent of Web 2.0 era, the Internet has become the world's largest repository of information, so Internet users constitute the Internet sources of information. As a typical representative in Web 2.0 era, micro-blog has obtained rapid development and wide application recently. However large user groups have produced a huge amount of information at the same time which is a challenge for information processing and research. At this point how to provide personalized service for users to filter high quality content and reduce the cost that user accesses to useful information effectively becomes very important [1]. The accurate entity relation extraction based on user interest is the prerequisite to realize personalized service.

However, existing entity relationship modeling on interest discovering are mostly based on micro-blog content and describe user's interests by extracting keywords from the text that rarely take advantage of the tag information in micro-blog [2]. Tags are user's description provided by micro-blog and the basis of recommendation for friend and information. Many micro-blog users marked their labels in order to show their own personal interests or attributes. In other areas, the studies about Internet have already some works to start thinking about using the tag information to represent user's interests. But in terms of micro-blog, a few people use tag to discover user interest and the existing methods based on tag do not consider interactive information between users fully [3]. So this article will take into account the personalized tag information and interactive information of user to conduct interest mining.

ISSN: 1975-0080 IJMUE

As a whole, the article analyzes the characteristics of micro-blog tag and intends to use the tag vector to describe user's interests. Then we will regard the problem of interest discovering as a tag recommendation for micro-blog user. After statistics we found that most of the tags are only marked by small number of users, so establishing user tag matrix in this way may produce serious problem of sparse matrix which can't accomplish tag recommendation effectively. Meanwhile, considering the influence that words of micro-blog user exert on recommendation, we build User-Tag and Tag-Word bipartite graphs respectively to solve the above problems [4]. Finally, in order to explore the relations between user's interactive behavior and tag, this paper produces a tag diagram by using the interactive relationship diagram of micro-blog user and puts forward a Tag Rank method based on PageRank thought [5], then sorts tags by calculating the importance of them so as to realize interest discovering of micro-blog user.

The rest of the paper is organized as follows. Section 2 discusses the related works. Section 3 analyzes the characteristics of the tag and illustrates our proposed methodology in detail. The extensive experimental results on the data set crawled from micro-blog website are presented in Section 4. We conclude our work in Section 5.

2. Related Work

The traditional methods of user interest modeling may not be fit for discovering micro-blog user's interest. For this reason the researchers conducted a lot of researches and explorations about micro-blog. In order to solve the problem of excessive noise data in micro-blog, Liu *et al*. [6] combined the methods of statistical machine translation and keyword extraction based on TFIDF to extract the keywords from the user's micro-blog describing their interests. Qiu Yunfei *et al*. [7] proposed an approach on micro-blog short text reconstruction to overcome the lack of information in short text after analyzing the structure and content of micro-blog short text. Chen [8] carried on the experiment about discovering user interests by using the micro-blog user's content and user's fans respectively. Wen [9] collected the micro-blogs issued by user and integrated them into one big file and then detected their potential interests of topic by using the LDA model. However, the methods mentioned above are mostly based on user's contents, focusing on the study and research of user browsing content, but the analysis of the user's behavior is still inadequate, failed to use user's behavior to discover their interests and filter the noise effectively. So the objectivity and accuracy of user interest model constructed by these approaches are limited.

Michelson [10] used entity of wikipedia category labels, established topic model under this classification for each user to determine what interests they were belong to. Zhao *et al*. [11] found that after comparing with traditional online media content by using topic model, people tended to talk about topics relating to family and life on Twitter. Ramage *et al*. [12] used Labeled - LDA for Twitter content, user modeling, the sort of micro-blog and user recommendation tasks. Literature [13] proposed a framework of user modeling based on Twitter and applied it to recommendation task further using the traditional media news and Hashtag in micro-blog. Literature [14] proposed a user model application TUMS based on Twitter. Given a twitter user with a collection of all the micro-blogs issued, riches them semantics, returned the result of user modeling and made it visualization. [15] aimed for a single micro-blog classification. The specific way was extracting the entities from the micro-blog, getting Wikipedia category nodes that correspond with each entity. And the node had hierarchy, thus we can obtain the category that each micro-blog belongs to through the algorithm based on the path. Literature [16] proposed a method based on a set of user browsing behaviors and established a calculation function of interest degree with variables browsing rate, stay time and browsing speed of webpage. However the feature of representing methods for user and document was single, they usually described the document or the user as a "whole"

leading to the characteristic representation lacking of hierarchy, limiting granularity of expression and reflecting the purpose of "personalized" difficultly.

3. The Model of the Approach

In this section, we mainly illustrate some related knowledge with regard to the tag, including characteristics and some problems existing in tag recommendation. Then modeling algorithm based on bipartite graph and PageRank thought is discussed in detail in the second subsection. At last the framework of the interest model is proposed.

3.1. User-Tag Bipartite Graph

Social tagging system allows user to use any words marking resources of their own or someone else. Compared to manual tagging system, social tags are data resources which users marked by themselves. Due to the huge amount of data, the quality of marking tag by user is less than professional, as long as we make the full analysis and use of these data, it will have tremendous value for data mining. So the paper intends to use the tag information representing user interest and put forward a method based on bipartite graph for interest discovering. User-Tag bipartite graph represents the initial connection relationship between users and tags which is shown in Figure.1. In the traditional bipartite graph model, if the user is assigned a tag, then the user has a side connected with this tag, the weight of edge and that of initial tag is 1.

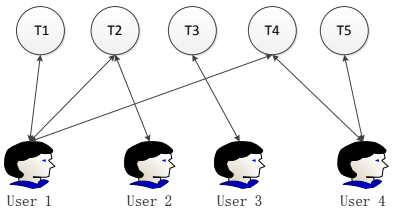


Figure 1. User-Tag bipartite Graph

However it will generate two problems: first bipartite network is too sparse. We perform a statistical analysis of user distribution with different number of tag on a 2 million datasets about micro-blog users. The result is shown in Figure.2. The vast majority of users do not mark the tags which are more than half of the total number of users. Meanwhile, the number of users labeling 10 tags is significantly higher than the number of users labeling 1-9 tags, even up to 2-3 times. This may be due to number limitation that tags are allowed to be marked up to 10 for micro-blog, so the users tend to mark the full tags. The phenomenon in Figure shows that many tags could not be made secondary distribution in the traditional bipartite graph model. For those users who did not assign tags for themselves, there will be no side connected with a tag which means they cannot use bipartite graph algorithm to make the allocation of tag resources. Second, average allocation method does not take into account the user preference for the tag. Due to the weight of tag assigned by user is defined as one, this will do not show user genuine interest directly.

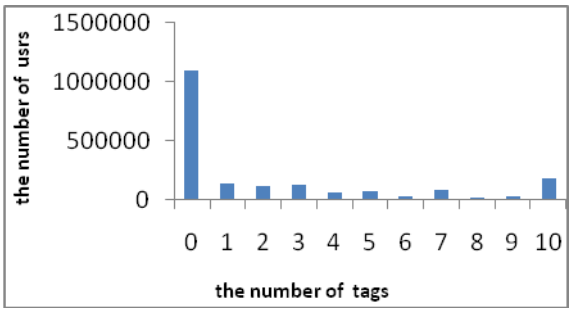


Figure 2. The User Distribution of Different Numbers of Tags

Because the micro-blog posted by user also shows his interest, so the paper considers using micro-blog information to solve the two problems. Obviously, if a tag appears frequently in the user's micro-blog, this means the user contains the interest described by this tag. And the occurrence number of the tag shows the user's preference to this tag. So consider using the occurrence number of tag in user micro-blog to describe the weight of User-Tag bipartite graph.

Weighted User-Tag bipartite graph constructor method is as follows:

U represents user set, T represents tag set, E represents link relationship between user and tag. $\{u_1,u_2,u_3,\ldots,u_n\},\{t_1,t_2,t_3,\ldots,t_m\}$ represent the node set of U and T respectively, E_{ij} represents the weight between user i and tag j, weight is set as the occurrence number of tag in micro-blog.

We can get the weighted single-mode mapping graph G of tag set T by the weighted bipartite graph.

The weight formula between nodes in G is defined as below:

$$W_{ij} = \frac{1}{K(t_j)} * \sum_{l \in U} \frac{E_{li} * E_{lj}}{K(u_l)} \quad (1)$$

$K(t_j)$ represents the degree of tag j in User-Tag bipartite graph, the formula is:

$$K(t_j) = \sum_{l \in U} E_{lj} \quad (2)$$

$K(u_l)$ represents the degree of user l in User-Tag bipartite graph, the formula is:

$$K(u_l) = \sum_{j \in T} E_{lj} \quad (3)$$

W_{ij} represents the percentage of the value of tag j assigned to tag i.

For user i and tag j, the initial value of j is $f(j) = E_{ij}$, the calculation formula for final recommendation value of j is shown below :

$$f'(j) = \sum_{a \in T} W_{ja} * f(a) \quad (4)$$

Set the initial tag vector of user is $R=\{f(1),f(2),\ldots,f(m)\}$, so the final vector is:

$$R' = W \times R \quad (5)$$

3.2. Tag-Word Bipartite Graph

Take advantage of User –Tag bipartite graph is based on the use of other users who have the same tag and assign the tag to user. The result obtained by this way is a coarse-grained result which has some disadvantages. First, the tag of common type is used by a large number of users leading to the tag will be assigned higher weight. But the weight of personalized tag that has more resolution for user will be smaller. Second, we do not have a resolution for the polysemy of tag. For example, apple can express both fruit and phone. For those users who use this tag containing different meanings, the relationship will be made between users based on User-Tag binary graph to distribute tag, but there is no relation between them in fact. The reason is that we do not consider the micro-blog information produced by user. If introduce the connection between tag and micro-blog word then we can eliminate the effects of the two problems to a certain extent. For the

first question, plus the value of tag that associated with word in user micro-blog, so the ranking of that personalized tag will increase. For the second question, if the word is related to food, nutrition in user micro-blog then we can assert that apple expresses fruit. So the recommended tag should be associated with the fruit. If the word is related to product, communication then we can assert apple expresses phone. So the recommended tag should be associated with the phone. Therefore on the basis of the result, the paper tries to use Tag-Word bipartite graph correcting tag information to get more precision result.

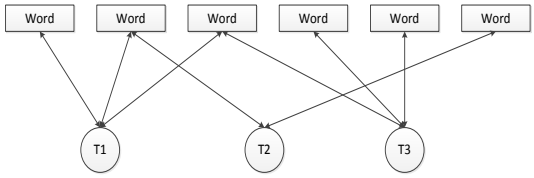


Figure 3. Tag-Word Bipartite Graph

In Tag-Word bipartite graph, T represents tag set, C represents word set, E represents the link relationship between tag and word. $\{c_1,c_2,c_3,\ldots,c_n\},\{t_1,t_2,t_3,\ldots,t_m\}$ represent the node set of C and T respectively, E_{ij} represents the weight between c_i and t_j , weight is set as the co-occurrence number of c_i and t_j in micro-blog. So we can get the weighted single-mode mapping graph G' of tag set T.

The weight formula between nodes in G' is defined as below:

$$W_{ij}' = \frac{1}{K(t_j)} * \sum_{l \in C} \frac{E_{li} * E_{lj}}{K(c_l)} \tag{6}$$

$K(t_j)$ represents the degree of tag j in Tag-Word bipartite graph, $K(c_l)$ represents the degree of word l in Tag-Word bipartite graph. W_{ij} represents the percentage of the value of tag j assigned to tag i.

For word w and tag j, the initial value of j $g(j)=f'(j)$, the calculation formula for final recommendation value of j is shown below :

$$g'(j) = \sum_{a \in T} W_{ja}' * g(a) \tag{7}$$

At this time the tag vector formula of user is:

$$R'' = W' \times R' = W' \times W \times R \tag{8}$$

Finally, we need to plus the influence that word in user micro-blog made on tag recommendation, assume that the word vector of user is H, so the weighting tag i is:

$$h(i) = \sum_{l \in C} \frac{H_l * E_{li}}{K(c_l)} \tag{9}$$

$K(c_l)$ represents the degree of word l in Tag-Word bipartite graph, the weighting tag vector is defined as R'''.

$$R''' = H \times E \tag{10}$$

The final tag vector of user is:

$$\vec{T}_1 = \alpha R'' + (1-\alpha) R''' \ (1 > \alpha > 0) \tag{11}$$

α is a harmonic factor that adjusts the influence of user's word vector on the tag. The greater the α, the smaller the influence of user's word vector is.

3.3. User Interest Modeling Based on Weighted Bipartite Graph

The general process of recommendation algorithm of weighted bipartite graph is follows: First establish the initial matrix of weighted bipartite graph. Second, establish a single-mode mapping matrix of a set in bipartite graph by initial matrix of bipartite graph. Finally, make the resource redistribution on the initial matrix of bipartite graph by using the single-mode mapping matrix to get the final matrix of bipartite graph.

User interest discovering algorithm based on bipartite graph gets the final result through the establishment of User-Tag and Tag-Word bipartite graph. The description of the algorithm is as follows:

Algorithm 1: The algorithm of discovering user interest based on weighted bipartite graph

Input : UserSet, TagSet, UserMbSet, UserTagM, WordSet and α
UserSet, TagSet, UserMbSet and WordSet is a set of user, tag, user micro-blog and word respectively. UserTagM is a matrix of user tag. α is a conditional factor of word vector influence.

Output: $\vec{T}_1$

1 // build User-Tag bipartite graph $M_{u,t}$, the initial matrix of bipartite graph is initial matrix of user tag
2 $M_{u,t}$ = UserTagM
3 for each user u in UserSet
4 // determine whether there is a tag for each user micro-blog
5 for each micro-blog in UserMbSet
6 if Tag t in micro-blog
7 $M_{u,t}$ ++
8 end for
9 end for
10 establish a single-mode mapping matrix G of tag set by $M_{u,t}$ and get the weight W between the nodes in G by formula(1)
11 make a redistribution for the weight of tag by G and get tag vector $R^{'}$ of user by formula(5)
12 // establish User-Word vector H and Tag-Word bipartite graph $M_{t,w}$
13 for each user u in UserSet
14 for each micro-blog in UserMbSet
15 if word w in micro-blog
16 $H_{u,w}$++
17 if tag t in micro-blog
18 $M_{t,w}$++
19 end for
20 end for
21 establish a single-mode mapping matrix $G^{'}$ of tag set by $M_{t,w}$ and get the weight $W^{'}$ between the nodes in $G^{'}$ by formula (6)
22 make a redistribution for the weight of tag by $G^{'}$ and get tag vector $R^{''}$ of user by formula(8)
23 // get the weighting result by using the word in user micro-blog
24 for each user u in UserSet
25 for each word w in H_u
26 for each tag t in G_t
27 $R^{'''}_{u,t} = \dfrac{H_{u,w} * G_{t,w}}{\text{totalnum of w in G}}$
28 end for
29 end for
30 end for
31 $\vec{T}_1 = \alpha R^{''} + (1-\alpha) R^{'''} \quad (1 > \alpha > 0)$

3.4. The TagRank Algorithm Based on PageRank

Human behavior is highly correlated with their own interests. Lv En psychologist suggested that human behavior is a function of a human body and the environment based on force field theory; Psychologist Xi Erjia proposed that internal driving force theory explains the phenomenon of biological control (User information behavior depends on the internal driving force and habit strength under certain conditions) [20]. For micro-blog user, their behavior can be understood as user interactive information including followers, forwarding and comments, which show the user's interests indirectly. Therefore we need to collect users that somebody pay attention to, forward to and comment on as a collection of his interactive relation.

The algorithm needs to build a tag relational graph firstly. Define the Ui as a user set who interact with user i, and then define the Ti as a tag set that users in Ui marked. Create a graph G (V, E) where V stands for tag nodes and E for all sides. When two tags appear in the same user, then there is an edge between the two tags linked each other. W_{ab} stands for co-occurrence frequency of tag a and b.

Sorting algorithm based on graph uses graph structure to determine the importance of node essentially. The basic idea is that every time a node is connected to another node, then the node will throw a vote for a node that it linked to, the more votes a node get the more important the node is. For a tag relational graph, the relationship between the tags that have more co-occurrence frequency is much more closely than those that have less co-occurrence frequency. The importance value of the tag will be assigned according to the weight of co-occurrence frequency which is different from PageRank algorithm with the importance value of the node divided equally. For example, in Figure.4 the node A will assign the equal weight to B, C, D in PageRank. While the node Ta will assign the weight of 1/5, 3/10, 1/2 to Tb, Tc, Td by the percentage of co-occurrence frequency.

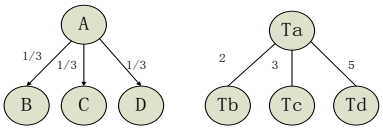


Figure 4. Weight Distribution Contrast

As we know that calculation formula of PageRank is as follows, L_j represents the number of outlinks of point j.

$$P_{l+1}(V_i) = (1+d) + d * \sum_{j \in E_{ij}} \frac{1}{L_j} * P_l(V_j) \tag{12}$$

Therefore, the calculation formula of TagRank algorithm in this paper is defined as follows:

$$P_{l+1}(V_i) = (1+d) + d * \sum_{j \in E_{ij}} \frac{w_{ij}}{\sum_{k \in Ejk} w_{jk}} * P_l(V_j) \tag{13}$$

w_{ij} represents the co-occurrence frequency of tag i and j, d is a random probability that represents the probability from one node to another node, defined as 0.85. In conclusion the flow chart of algorithm is shown below:

Figure 5. Flow chart of TagRank algorithm

The description of the algorithm is as follows:
Input: the training set of micro-blog user
Output: the interest vector of micro-blog user

Step 1: Collect interactive users and build a set M of user interactive relationship;

Step 2: Collect tags from M and build a tag diagram D by the co-occurrence relationship between tags;

Step 3: Calculate the weight of each tag in D according to formula (13);

Step 4: The procedure will not end until the weight of each tag do not change;

Step 5: Sort tags by weight and select the K tags with highest weight as interest vector $\vec{T_2} = \{t_1, t_2, t_3, ...t_k\}$;

The results obtained by above two methods have different emphases, so in order to gain the final user's interest, the results of these two algorithms need to get a comprehensive. Therefore use a combination of these two results by adopting a harmonic factor γ to get the final description of user interest vector. Define $\vec{T_1}$ as a vector of user interest obtained by the method based on bipartite graph, and define $\vec{T_2}$ as a vector of user interest obtained by the method based on TagRank, so the final user interest vector is calculated as:

$$\mathrm{T} = \gamma\vec{T_1} + (1-\gamma)\vec{T_2} \tag{14}$$

The overall framework of the approach is as shown in Figure.6:

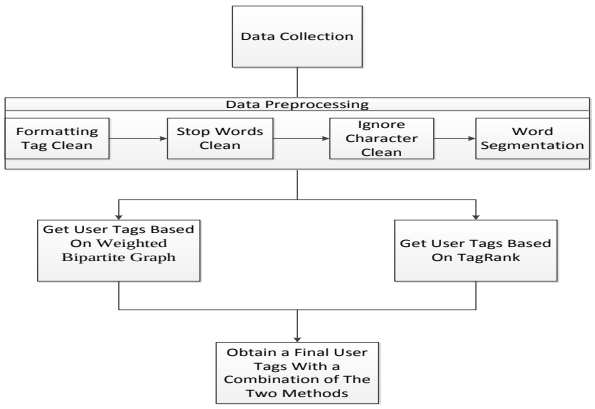


Figure 6. The Algorithm Framework of User Interest Modeling

This article is to solve the problem of user interest discovering in micro-blog, therefore, the data set is collected from the most popular Sina micro-blog in China. The collected data cannot be used for user interest discovering directly, so we need data preprocessing first. Preprocessing mainly contains the following: formatting tag clean, stop words clean, ignore character clean and word segmentation. The micro-blog content contains a variety of symbols with special features, such as "#", "[]", "//". These symbols are of no use for text data processing that need to be removed. In text set there are usually some function words, adverb, conjunction, preposition and so on. If we do not remove them, it will greatly increase the dimensions of the vector space model which is harmful for text processing. Ignore character clean means that text processing system does not require some characters, such as punctuation character, the Arabic numerals, web links and so on which need to be clean up. Finally, because there do not have some marks to represent interval between words in Chinese, so in order to extract word from Chinese text, word segmentation operation is required.

4. Experiment

In this section, the performance evaluation of our mechanism is conducted in order to verify the effectiveness of our proposed approach above. The simulation environment is

performed on Microsoft Windows 7 operating systems, AMD-A63420M 4-core CPU processor, 4 G memory, 500 G hard disk, and My Eclipse 10 software tool. For text preprocessing part, using the segmentation system of Chinese academy of sciences to conduct word segmentation and remove stop words. Our designed experiments mainly include four parts: conforming of α and γ, test results of the proposed approach and making comparisons with other algorithms. Experimental data come from Sina micro-blog API and select information of 500 micro-blog users including both VIP and ordinary user. The information include user's personal tag, the latest 500 micro-blog contents, forwarding contents, comments and grab followers information at the same time.

The experiment uses the comparison between user interest obtained by experiments and that by artificial observation way, then assesses the quality of interest model through the three indicators: Recall, Precision, and the F-value. According to the characteristics of the model representation, the calculation formulas are defined as follows:

$$\mathrm{Re}call = \frac{\text{the correct number of interests in model}}{\text{the actual number of users's interests}} \tag{15}$$

$$\mathrm{Pr}ecision = \frac{\text{the correct number of interests in model}}{\text{the total number of interests in model}} \tag{16}$$

$$F = \frac{\mathrm{Re}call * \mathrm{Pr}ecision * 2}{\mathrm{Re}call + \mathrm{Pr}ecision} \tag{17}$$

The accuracy of user interest model depends on Precision, the higher the Precision, the less the number of error interests; the number of related interests that missed by model depends on Recall, the higher the Recall, the less the number of effective interests that missed. F is a harmonic mean of the Precision and Recall which integrate the Precision and the Recall into one index, thus we can evaluate the quality of the system better. Generally speaking the higher the F-value, the better the quality is.

Experiment 1. The dynamic test results of α

Experiment firstly studies the influence of the harmonic factor α on result based on weighted bipartite graph, confirming α under what circumstances have the best result by F-value. The number of tag in model is set as10, the value of α is set to 0-1, step length is 0.05. The result is shown in Figure.7, when the value of α reaches 0.05, the effect of algorithm based on weighted bipartite graph is the best. So we set the value of α as 0.05 in next experiments.

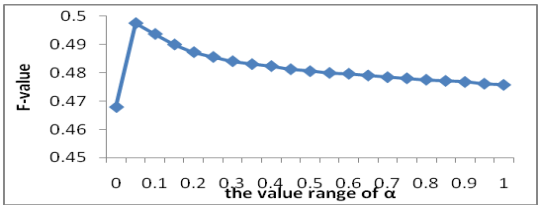


Figure 7.The Dynamic Test Results of α

Experiment 2. The dynamic test results of γ

Then we studies the influence of the harmonic factor γ on result, confirming γ under what circumstances have the best result through F-value. To determine the influence of the harmonic factor γ on the result, the γ is set to 0-1, step length is 0.1, the number of tag in model is set as10. The higher the γ, the higher the proportion of the modeling algorithm based on bipartite graph. The result is shown in Fig. 8, the effect of the proposed algorithm is less than that based on bipartite graph when the γ is smaller which means that the result is ineffective using the TagRank algorithm alone. It can be seen that after the combination of the two methods, F value is higher than using either method alone. Start with the increase of the γ, recommendation effect is better and better, and when the γ

reaches 0.3, the F-value achieves the max value, and then began to decline. So we set the γ as 0.3.

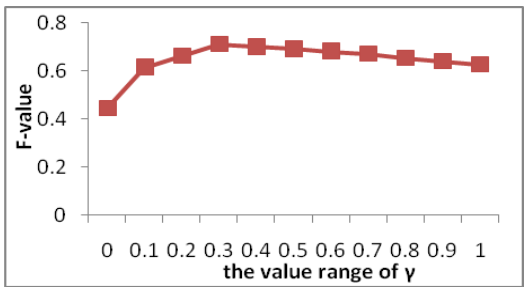


Figure 8.The Dynamic Test Results of γ

Experiment 3. The result of proposed approach

In order to investigate the utility of the model comprehensively, this article selects four types of the user from the data set. The no.1 is a expressive user who posts content frequently, the no.2 is a watching user who has little tags marked and post rarely, the no.3 is a social user
who has much interaction with other users and the no.4 is a comprehensive user who is active in micro-blog. The user interests (10 keywords) obtained by experimental model and that by artificial collecting for comparison is shown in Table 1:

Table 1. The Experimental Results Contrast

Serial Number	The Interests Obtained by Experiment	The Interests Obtained by User	The Interests Supplemented by User
1	optimism, dessert, singer, blue, travel, sing, network, collecting, writing, Guiyang	optimism, dessert, blue, travel, sing, network, colleting, Guiyang, writing	music fan, badminton
2	sports news, media, fitness, newspaper, television, advertisement, network, planetarium, recording, taxi	sports news, media, fitness, newspaper, network, planetarium, recording	basketball, camera, football
3	fashion, tide, delicious food, scenery, lose weight, fairy tales, quotations, dog, constellation, peach	fashion, tide, delicious food, scenery, lose weight, quotations, dog, constellation	badminton, running
4	scientific research, running, bicycle, English, speech, literature, Starbucks, smoking, drama, music	running, bicycle, English, speech, literature, Starbucks, drama	presiding, public benefit, current affair, classical music

According to the results in the Table 1, calculate the recall rate, precision and F-value, the data as shown in Table 2:

Table 2. The Experimental Results

Serial Number	Recall	Precision	F
1	0.82	0.90	0.86

2	0.70	0.70	0.70
3	0.80	0.80	0.80
4	0.64	0.70	0.67
Average value	0.740	0.775	0.758

According to the results in Table 2, experimental results are relatively ideal. Performance is better in judgment and identification for user interest model. By analyzing testing data of four different user types, the results are different between them. The cause of such difference is due to the different quality of the user data. In the Table 2, the maximal accuracy is 0.90, namely in 10 words describing user's interests, there is only one cognitive deviation word for user. And the test results tend to be stable on the whole. In the description of user's interest preference the model has a certain value.

Experiment 4. The performance contrast

We evaluate the performance of our proposed approach (OPA) compared with the traditional collaborative filtering [21] algorithm (CF) and TFIDF algorithm from the three indicators: Recall, Precision, and F-value. The principle of collaborative filtering recommendation algorithm is to recommend project that user who has similar interest likes. The TFIDF algorithm is extracting the effective keywords from the user micro-blog content. Because the number of tags in user's model will bring great influence for the final result, so experiment compares the effect of each algorithm under the circumstances with different number of tags. The results are shown in Figure.9-11.

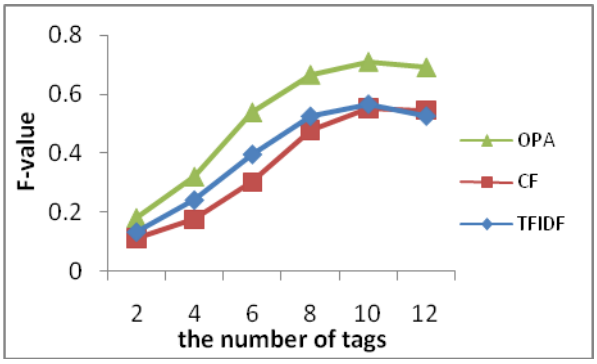


Figure 9. F-Value Contrast

As can be seen from the Fig.9, the F-value of three approaches is basically the same when the number of tags is less. With the number of tags increasing, the superiority of the approach in this paper is more and more obvious. No matter how much the number of tags, the effect of the approach in this paper is best. When the number of tags increases to ten, the F-value comes to maximum. In other words, when the number of tags increases to ten, the performance of the approach is best which gives consideration to the precision and recall rate. So the number of tags in interest model is inclined to be set as ten in order to discover user interest more effectively. When the number of tags is ten, the F-value of three approaches is respectively 0.711, 0.553 and 0.566. Compared with other two approaches, the F-value of the proposed approach increases by 28.57% and 25.62%.

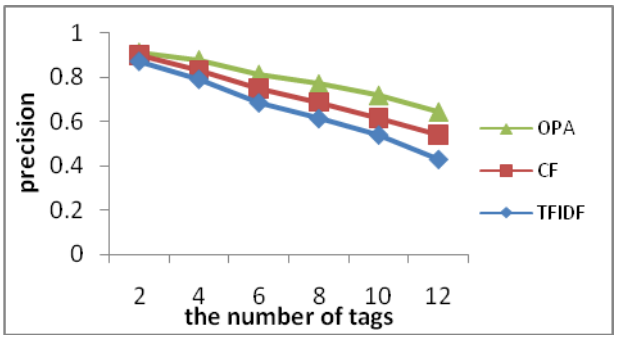


Figure 10. Precision Contrast

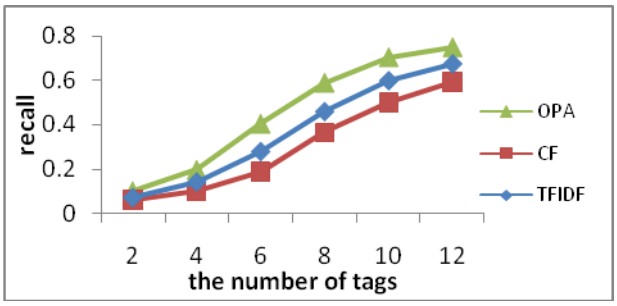


Figure 11. Recall Rate Contrast

As can be seen from Figure.10 and 11, with the number of tags increasing, the recall rate of user interest model rises but the precision decreases. This is because that the approach only needs to recommend a few tags when the number of tags in model is less which leads to the precision greatly increasing; But the less number of tags in model, however, leads to a large number of correct user's interests cannot be recommended, so the recall rate will be lower. With the increase of the number of tags, the wrong tags are rising and the proportion of correct tags that missed is falling. Therefore, when we set the number of tags in the model, we need to select the optimal number of tags with precision and recall rate considered. In addition, in the experiment CF method used collaborative thought to carry out tag recommendation that tend to be higher accuracy. While TFIDF method discovers user's interest based on keyword extraction that tends to be higher recall rate. So the two approaches focus on different aspects and the proposed approach combines both advantages leading to the performance improved.

In Figure.10, the average precision of three approaches is respectively 0.789, 0.721 and 0.653. Compared to other two approaches, the precision of approach proposed in this paper increases by 9.43% and 20.83% on average. In Figure.11, when the number of tags in model is less, the recall rates of three approaches are all lower which has no significance to compare to. As can be seen, the maximal recall rate of three approached is respectively 0.749, 0.594 and 0.674, the recall rate of the proposed approach increases by 26.09% and 11.13%. It can be seen that the approach proposed in this paper shows better performance in all aspects.

5. Conclusion

In this paper, a micro-blog user interest modeling based on tag is proposed to improve the performance for discovering interest. Due to much noise data in micro-blog, the paper comes up with using tag vector to describe user interest. First, the approach improves the existing bipartite graph model and puts forward interest discovering method based on weighted bipartite graph. The method uses the occurrence number of tag in micro-blog to

represent tag's weight and gets the User-Tag, Tag-Word bipartite graph. On this basis, obtains the user interest vector by weighting the information of user's word. Second, considering that interactive relationship between users also describes user interest, this article produces label relationship diagram by user interactive relation and then obtains the user interest vector by TagRank algorithm. Finally, integrate the results of the two algorithms to get the final user interest vector. We conduct extensive experiments to test the proposed approach, and the experimental results reveal that proposed approach performs well in all aspects.

Due to the resolving ability of each tag are different, the personalized tags that have more resolving ability will have weak weight because the number of people used them and frequency marked them are less. So how to determine such tag is the focus of the next step work.

Acknowledgements

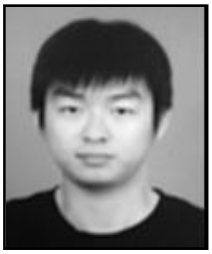
The work was supported by the National Natural Science Foundation (NSF) under grants (No.61472294, No.61171075), the Opening Project of State Key Laboratory of Digital Publishing Technology, Key Natural Science Foundation of Hubei Province (No. 2014CFA050), Applied Basic Research Project of WuHan, Program for the High-end Talents of Hubei Province, the Fundamental Research Funds for the Central Universities(WUT:2014-145210010). Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

References

[1] Y. S. Sun, W. Liu, R. R. Qiu and C. H. Huang, "Research Development of User Interest Modeling in China," Journal of Intelligence, vol. 32 no. 5, **(2013),** pp.145-149.
[2] M. Zhao, J. D. Song and H. H. E, "Interest Modeling Method for Web Users Based on Social Tagging Systems," Software, vol. 34 no. 12, **(2013),** pp. 137-138.
[3] H. Y. Min and S. L. Tu, "Ontology Based User model for Personalized Agriculture Search," 2012 IEEE Symposium on robotics and Application, Kuala, Lumpur, June 3-5, **(2012)**.
[4] "Bipartite network projection and personal recommendation," Physical Review E-PHYS REV E, vol. 4 no. 76, **(2007),** pp. 046115.
[5] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, vol. 30 no. 1, **(1998),** pp. 107-117.
[6] Z. Liu, X. Chen and M. Sun, "Mining the interests of Chinese microbloggers via keyword extraction," Frontiers of Computer Science, vol. 6 no. 1, **(2012),** pp. 76-87.
[7] Y. F. Qiu, L. Y. Wang, L. S. Shao and H. M. Guo, "User interest modeling approach based on short text of micro-blog," Computer Engineering, vol. 40 no. 2, **(2014),** pp. 276-279.
[8] J. Chen, R. Nairn, L. Nelson and M. S. Bernstein, "Short and tweet: experiments on recommending content from information streams," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, April 10-15, **(2010)**.
[9] J. S. Weng, E. P. Lim, J. Jiang and Q. He, "TwitterRank: finding topic-sensitive influential twitters," Proceedings of the third ACM international conference on Web search and data mining, New York, USA, February 5, **(2010)**.
[10] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: A first look," Proceeding of the Workshop on Analytics for noisy unstructured text data, New York, USA, October, **(2010)**.
[11] W. X. Zhao, J. Jiang, J. S. Weng, J. He, E. Lim, H. F. Yan and X. M. Li, "Comparing twitter and traditional media using topic models," Proceedings of the 33rd European Conference on IR Research, ECIR, Dublin, Ireland, April 18-21, **(2011)**.
[12] D. Ramage, S. T. Dumais and D. J. Liebling, "Characterizing Microblogs with Topic Models," ICWSM, vol. 10, **(2010),** p. 1.
[13] F. Abel, Q. Gao, G. Houben and K. Tao, "Analyzing User Modeling on Twitter For Personalized News Recommendations," Proceedings of the 19th International Conference, UMAP, Girona, Spain, July 11-15, **(2011)**.
[14] T. Tao, F. Abel, Q. Gao and G. Houben, "TUMS: Twitter-based User Modeling Service," Proceedings of the 8th international conference on The Semantic Web. ESWC, Heraklion, Greece, May 29-30, **(2011)**.

[15] Y. Genc, Y. Sakamoto and J. V. Nickerson, "Discovering context: classifying tweets through a semantic transform based on Wikipedia," Proceedings of 6th International Conference, FAC, Orlando, FL, USA, July 9-14, **(2011)**.
[16] Y. G. Xia and Y. H. Liu, "An improved method to calculate user's interest degree and amend user's interest," Journal of Modern Information, vol. 34 no. 1, **(2014),** pp. 46-48.
[17] W. Wu, B. Zhang and M. Ostendorf, "Automatic generation of personalized annotation tags for twitter users," The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, USA, June 2-4, **(2010)**.
[18] T. Lappas, K. Punera and T. Sarlos, "Mining tags using social endorsement networks," Proceeding of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, Beijing, China, July 25-29, **(2011)**.
[19] Y. Yamaguchi, T. Amagasa and H. Kitagawa, "Tag-based user topic discovery using twitter lists," 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Kaohsiung, Taiwan, July 25-27, **(2011)**.
[20] G. Y. Fu and Z. Y. Zhu, "User interest modeling based on the analysis of the behavior in personalized service," Engineering and Computer Science, vol. 27 no. 12, **(2005),** pp. 76-78.
[21] S. R. Chen, "User interest modeling and application research for personalized service," Chongqing: Chongqing University, **(2007)**.

Authors



Junjiang Li, He is a M.S. student in the Department of Computer Science at Wuhan University of Technology. He received his B.S. degree in Software Engineering from Zhengzhou University in 2011. His research interests are in cloud computing.

Chunlin Li, She is a Professor of Computer Science at Wuhan University of Technology. She received her M.S. in Computer Science from Wuhan Transportation University in 2000 and her Ph.D. in Computer Software and Theory from Huazhong University of Science and Technology in 2003. Her research interests include cloud computing and distributed computing.

Youlong Luo, He is a vice Professor of Management at Wuhan University of Technology. He received his M.S. in Telecommunication and System from Wuhan University of Technology in 2003 and his Ph.D. in Finance from Wuhan University of Technology in 2012. His research interests include cloud computing and electronic commerce.

