A friend recommendation algorithm based on the user relationship

Qi SHEN[1, a], Sibo WANG[1, b], Ran WANG[1,c] and Ke CAO[1,d]

[1] Department of Software Engineering, Beijing University of Technology, Beijing, 100124, China

[a]email:shenq@bjut.edu.cn, [b]email:314753529@qq.com, [c]email: 547066756@qq.com, [d]email: 790533531@qq.com

Keywords: Tags; User Relationship; Text Classification; Friend Recommendation.

Abstract. In recent years, the micro blog in the social networks in China is becoming more and more important, people use micro blog more frequent, people can follow to the people they want to know, and people who have common interests. Based on the analysis and research of the existing micro blog friend recommendation algorithm, this paper proposes a friend recommendation algorithm based on user relationship in micro blog, improve the accuracy of recommendation.

Introduction

Communication refers to the transmission of information between people and people, as well as between people and society. Common media are home phone, mobile phone, newspaper, TV and computer networks etc. Among them, the network spread in the world are Facebook, Twitter and so on, China has Renren, Sina micro blog, etc. After Sina micro blog officially launched in 2009, it was widely acclaimed. Sina micro blog 2015 active users increased to 210 million, growth of 35 million over the end of last year. It is one of several major features; users can easily post micro blog forwarding, comments. A micro blog length is usually less than 140 words. Micro blog can usually be the first time to reflect the user's personality and their hobbies. The other is to pay attention to their favorite stars and the state of interest

At present, Sina micro blog friend recommendation mechanism, according to the user information to fill in the educational background, custom tags, friends of friends and the user's location write in the user’s background and other ways to recommend. Users can also click to view more connections, See according to celebrity recommendation, expert recommendation, the official certification recommended for more detailed search. The recommended order is based on the number of fans, although there are some simple classifications, but did not according to the needs of different users, to make more personalized friend recommendation. In this era of big data, the use of the available information in the micro blog is still far from enough. So this paper proposes a comprehensive friend recommendation algorithm from many aspects.

Related works

If a user follow too many people, will probably not be able to find their most interesting micro blog from a large number of micro blog, so it is particularly important to recommend to the user that he might be interested user. At present the common recommendation algorithm has the content based recommendation algorithm and the collaborative filtering recommendation algorithm [1]. Content based recommendation algorithm[2], can be based on the user's personal information, tags, etc.. The collaborative filtering algorithm is divided into user based collaborative filtering, collaborative filtering based on item two.

Nan Zheng according to the micro blog time for friend recommendation [3], Yiwen Zhang through the user's custom tags to carry out a friend recommendation [4], Xiaoli Lin use Python language to the user's interest and the number of fans and the number of users micro blog is forwarded by the number of comments were clustering, in-depth mining of the relationship between the micro blog users.[5] Yang Changchun based on existing Sina micro blog friends recommend deficiencies, according to the influence of the authority of users to improve the friend

recommendation algorithm [6]. Guoyan Huang made the friend recommendation based on user interest [7].

This paper in order to make full use of micro blog users' information, use the relationship between the two users, from the user's Micro blog and the user's location, an improved recommendation algorithm is proposed.

A friend recommendation algorithm based on the user relationship

User relationship analysis

At present, many social networking sites friend recommendation common method is "friends of friends", that is, if user A and user B are friend, and user B and user C are also friend, then user A may be with the user C is also a potential friend. With the equation can be expressed as: Friend(a,c)={ Friend(a,b)∩Friend(b,c)}.

Micro blog and other social networking sites are different, It is a one-way follow social way. Because micro blog is not only looking for our recognition of the potential people, it also provides opportunities for us to find people with similar interests. Although there are many Micro blog users plus V(V is the verification of celebrities). They have a lot of fans, but if only recommend the user who plus V with a lot of fans, and ignore the interest, the effect will be not good of the friend recommendation.

Two common social network analysis tools are compared in paper[8], Ucinet and Gephi, Ucinet is more suitable for dealing with complex problems of multiple relationships, so this paper uses Ucinet to deal with the relationship between the user and the user. The schematic diagram of the relationship between some micro blog users is shown in Figure 1. Can be seen from Figure 1, the user's attention between the micro blog is one-way, not a user follow a user, and another user will certainly follow him. Mining whom we follow each other users, it is more likely to find other users we are interested in.

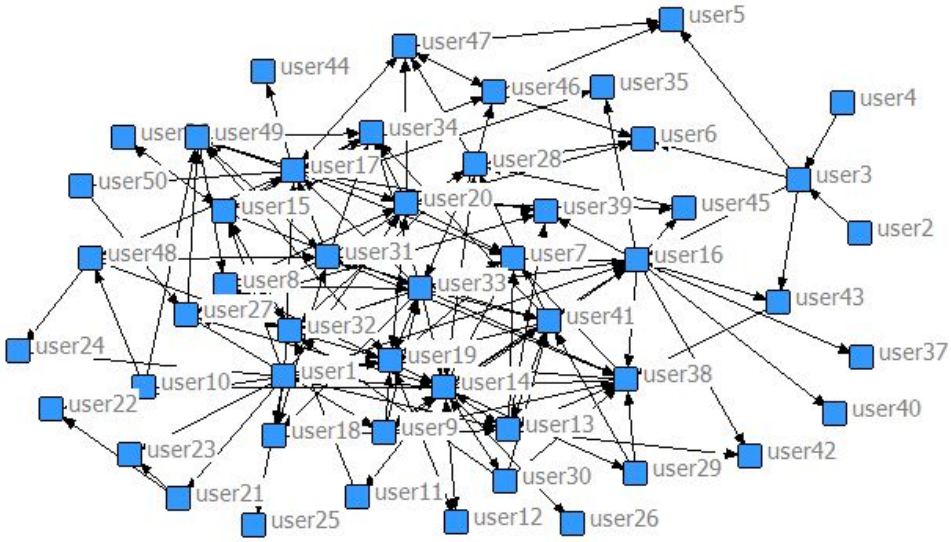

Fig. 1 The relationship between 50 users directed graph

Considering the effect of the relationship between users, for users more familiar with the user, their follow will be more weight. We will be record each users who reply from the user's sent Micro blog information, cumulative most of the replies, he is more likely to be the most closely related user of user many friends. For users send micro blog @ people, @ representative want the first time for him to see this message. The @ action will be recorded. As shown in the equation (1), is a friend of mutual follow with the user, the equation of the user side influence:

$$Infl(a,b)=0.1inter(a.b) \quad (1)$$

Influence refers to a person's action, may bring about the impact of the people around. inter(a.b)is the number of reply information's number and the @ number between user A and B,0.1 is not to let this part of the proportion of too much is set manually. N for the user A all users follow about each other, the equation for calculating a potential friend is:

$$\text{Innerfriend(a,c)}=\sum_1^n[friend(a,c)+infl(a,b)] \quad (2)$$

The use of personal information

When looking for the users we want to follow, we can start from the gender, in the same city, micro blog rank. The higher the rank of micro blog users, indicating that they usually use a higher frequency, may be more timely communicate with other users. At present the highest rank of micro blog users is 39, in this paper, the use of micro blog is more than 5 ranks of users.

The clustering analysis of user Tags

Because the tag itself is not very long, so do not have to go through the word segmentation tools, but many users do not use the tags, we count all the tags in the experiment, and choose the highest rate of 50, I use the people Xia Tian of the Renmin University of China developed Xsimilarity, divides the similarity between the tags.

Tag A, the similarity between B with Similar (a, b) said. If the similarity of two tags in 0.7 above, we put these two tags into one. Such as music, listen music, travel and tourism, here respectively merged into one tag. After selected using the largest number of 10, others using less tags we no longer use. Here we refer to the clustering method used in this paper[5]. Randomly selected 100 users, use at least any of the 10 Tags of a user. Statistics concerned about two kinds of user tags at the same time, constructed as shown in Table 1. For example, in the first row and the second column number 42, is the representative of the tag both the movie and music, the number of people is 42.

The data were clustered with K-means clustering method. We set the number of clusters of 3, that is, a total of ten tags into 3 groups. The results obtained are music and 80's as a group, movie, travel, fashion, news, emotion; game become a group, food, youth to become a group. That is, the same group of users who are more likely to have a common interest.

Table 1 k-means Clustering statistics

	music	movie	travel	food	fashion	news	80's	youth	emotion	game	group
Music	0	42	47	39	35	27	25	18	11	36	1
Movie	42	0	38	39	22	25	31	28	17	29	2
Travel	47	38	0	37	28	24	40	25	22	23	2
Food	39	39	37	0	36	28	28	22	19	32	3
Fashion	35	22	28	36	0	25	34	27	21	25	2
News	27	25	24	28	25	0	26	21	16	18	2
80's	25	31	40	28	34	26	0	34	25	31	1
Youth	18	28	25	22	27	21	34	0	24	26	3
Emotion	11	17	22	19	21	16	25	24	0	19	2
Game	36	29	23	32	25	18	31	26	19	0	2

The use of user location

Live in the same city, or the regular activities of position relatively close to the people, the more likely to interact in addition to the micro blog, maybe even in real life to know each other. This paper is based on Sina micro blog information for friend recommendation，where can get the user state location of latitude and longitude. The latitude and longitude obtained after the decimal point has 6 digits, according to the calculation equation of latitude and longitude, if two people are within 30 kilometers, is the daily activities of the relatively close range of people. They could potentially be friends.

Users post micro blog classification

For users to post their own or forward micro blog information, because the total length of a maximum of 140 words each, different from the custom tags is the need to information extraction. The method is used ICTCLAS developed by the Institute of computer science, Chinese Academy of

Sciences to word segmentation, and the use of Harbin Institute of Technology stop list. Then use our own written words library, mainly partial entertainment, sports, for example, the sports star first merged into the club, and then merge the club into the sports field. Its frequency is calculated by the TF(term frequency), select the top 100 largest proportion of vocabulary, and delete other words. Using KNN text classification method [9], as shown in the equation (3), divided into IT, entertainment, sports and other aspects. The j is the category, m_{aj} and m_{bj} are the degree of membership in category j of user a and user b.

$$S(a,b)=\frac{\sum_j m_{aj} m_{bj}}{\sqrt{\sum_j m_{aj}^2 * \sum_j m_{bj}^2}} \tag{3}$$

The final friend recommendation

First, the first step is based on the user rank, the number of followers and the number of fans to judge, micro blog is requirement of rank 5 and above, the number of follow users and the number of fans are more than 20 users. According to the above friends influence recommendation equation to select 50 users. If the user does not comply with the requirements, select 100 users direct use of the above methods to friend recommendation based on geographical location. Next no matter what the use of the two methods, select 20 users, both need to be through by the user's post information and user's tags, and user's follow users' post information and tags, to determine all the interests of a user. Compare with the user's interest, if the interest is relatively low, here is set for the common interest is less than 2. And have not previously experienced location screening, then rescreening it. If the interest of high similarity, direct the final friend recommendation. Interest similarity is low, and has been screened by geographical location, it is also directly recommended. Figure 2 for the whole process.

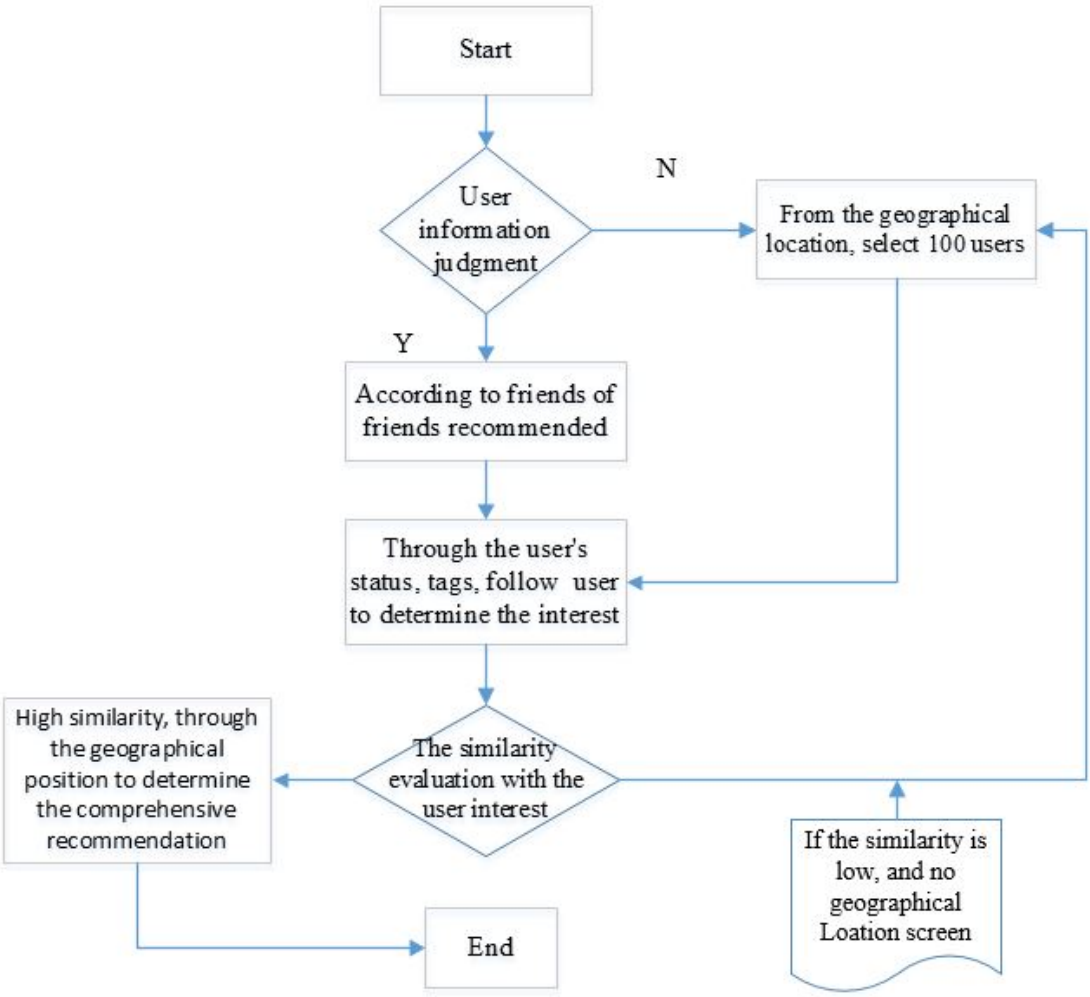

Fig. 2 The whole process

The experiment results and analysis

The experimental data acquisition

The data of this paper is obtained through API, which is provided by Sina micro blog open platform, since the interface was upgraded in 2014, some API, for example, user follow users' ID and user fans' ID, cannot get now. So this part of the data obtained from the website datatang. Finally, the actual experimental data obtained by combining two aspects. This paper is using the 63461 user information is from April 11, 2014 to May 11, 2014 of the data.

The analysis of experimental results

In this paper, precision is used to measure the experimental results. As shown in the equation (4). Friend recommendation by the improved algorithm, if the recommended user is already a friend of the user, then it is a precise recommendation, its symbol is recommend $_{correct}$ .All friend recommendation symbol are recommend $_{all}$. The greater the accuracy rate, the higher the accuracy of the algorithm. At the same time with the commonly used several recommended methods, based on the content of the recommendation, based on friends of friends recommendation, the two methods for comparison. The results are shown in figure 3. Take the recommended number respectively: 4,8,12,16,20, as a reference for accuracy. From the results we can know that the method proposed in this paper improves the accuracy of the recommendation to a certain extent.

$$\text{Precision}= \frac{recommend_{correct}}{recommend_{all}} \tag{4}$$

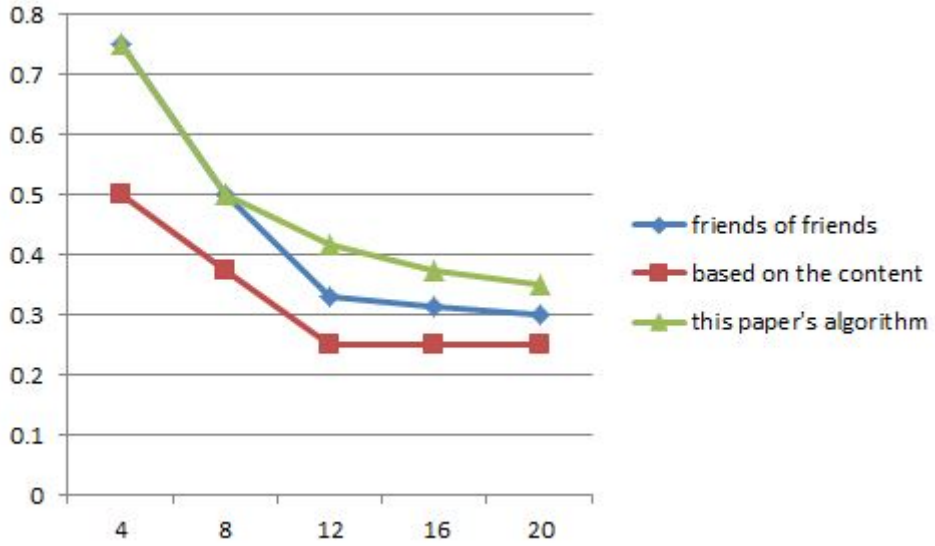

Fig. 3 Comparison of the effect of the three algorithms

Conclusion and future work

This paper analyzes the possibility of improving the accuracy of the user's friend recommendation by multi angle analysis; make full use of the role of the user relationship in the friend recommendation, proposes an improved friend recommendation algorithm, compared with previous studies, the accuracy of the previous research has improved. But because of the particularity of sina micro blog length limited, sometimes when people post important content, because of the number of words, usually in the form of pictures as a vector, in this paper, the picture does not have any extraction operation. At the same time, the text information is extracted from the user's post micro blog; the text library was set up manually, the content is not enough, in the follow-up study, we can further study and supplement.

References

[1], Goldberg D, Nichols D, Oki B M, et al. "Using collaborative filtering to weave an Information tapestry[J]". Communications of the ACM, 1992

[2],Fan Tang, Bofeng Zhang, Jianxing Zheng, Yajun Gu “Friend Recommendation Based on the Similarity of Micro-blog User Model” 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber,Physical and Social Computing

[3], Nan Zheng, Shuangyong Song, and Hongyun Bao “A Temporal-Topic Model for Friend Recommendations in Chinese Microblogging Systems” IEEE Transactions on Systems, man, and Cybernetics: Systems, VOL. 45, NO. 9, SEPTEMBER 2015

[4], Yiwen Zhang, Lihua Yue, Yifei Zhang, Qing Li, Jiaxing Cheng, “A friend recommendation method based on common users and similar Tags”. Journal of computer applications，2013，33(8) : 2273 － 2275

[5], Xiaoli Lin, Keke Hu, Qing Hu, “Research on the relationship of micro blog users based on Python” journal of intelligence Vol.33 No.6 June 2014

[6], Changchun Yang, Jing Yang, Hong Ding. “A new kind of sina micro blog friend recommendation algorithm”. Computer Applications and Software Vol. 31 No.7. Jul.2014

[7], Guoyan Huang, Yixin Chang, Yakun Li, “New Friend Recommendation with User Interest and Socialization” Journal of Information & Computational Science July 20, 2015.

[8]. Jun Deng，Xiaojun Ma，Qiang Bi ,“A comparative study of social network analysis tools Ucinet and Gephi”, Information Studies:Theory & Application Vol. 37 No.8. 2014

[9]. Xiulan Hao, Xiaopeng Tao, Hexiang Xu, Yunfa HU, “A solution to the problem of KNN text classifier deviation” Journal of Computer Research and Development 46(1):52-61, 2009