An improved PageRank algorithm for Social Network User's Influence research

Peng Wang, Xue Bo*, Huamin Yang, Shuangzi Sun, Songjiang Li

School of Computer Science and Technology,Changchun University of Science and Technology

Email:b8954471@126.com

Keywords: Social network; PageRank algorithm; User's influence; Forwarding rate; Activity degree

Abstract. In social networks, Micro Blog has become the most widely used social networking platform. Mining valuable customers is particularly important in the large-scale micro-blog user group, and the influence of the user become the main measure to estimate the user's value. Based on the similarity of micro-blog user groups and Internet node characteristics in structure, this paper proposed an improved PageRank algorithm by user forwarding rate and activity degree estimate the user's influence. Experiment results show that the improved algorithm has good convergence and the influence of the micro blog users can estimate effectively and objectively.

Introduction

Micro-blog as a new platform for social networking tools get more and more attention. In micro-blog often have such a phenomenon that some users forwarding or comments micro-blog will become a hot topic. These users have a high influence on the micro-blog network, they affect the dissemination of information, the trend of public opinion. The research on the influence of the user is beneficial to the discovery of the law of the spread of the micro blog and provides a theoretical basis for the advertising and public opinion management.

At present, there are some algorithms for the research of social network user influence. For example, Daniel M. Romero[1] who considered the influence of user influence with the fans and the number of fans, and thus proposed a HITS based algorithm to evaluate the influence of Twitter users. But this algorithm does not take into account the user's forward rate, it is easy to appear the situation when in estimating user influence fans are very positive but not forwarded. Heidemann[2] who proposed the link between users regarded as a weighted undirected edge, the weights between users are estimated by the user's interaction frequency in a certain period of time. Based on this weight to improve algorithm, and then estimate the influence of micro-blog users. This algorithm will be very complicated in the case of too many data.

In this paper, the forwarding rate and user's active degree of these two points as a measure of the impact of the user's influence. Forwarding rate is determined by the degree of attention among users, and attention is determined by the number of fans forwarding. The number of fans forwarded to prove a higher number of forwarding rate, the higher the user's forwarding rate, the higher the user's authority. With activity degree as the basis is because there are a lot of zombie users have a high number of fans, but they do not often publish news. Because the zombie users in the number of advantages, so it will lead to more authoritative than many new authoritative user. So this paper takes the user's activity as one of the indexes to estimate the user's influence. Considering the characteristics of micro-blog users. This paper based on the PageRank algorithm, by increasing the activity degree and forwarding rate to improved PageRank algorithm.

PageRank algorithm

PageRank algorithm is a calculation based on hyperlinks between pages ranking algorithm. PageRank algorithm is used to determine a page level through the network in countless hyperlinks, calculate the PageRank value for each page, and then sort web page based on the value. It defines a hyperlink from A pages to B pages as a vote of A pages to B pages, determine the new level by the

source of the vote and the rank of the target. Web page PageRank values can be obtained by the following formula:

$$PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \ldots\ldots\right) d + \frac{1-d}{N} \quad (1)$$

Where d is the damping coefficient . A, B, C, D, and so on behalf of the page. L is the number of out of chain pages. d after Page Larry several experiments show that the experimental results are best when the 0.85[3] is taken.

This paper improved PageRank algorithm to study the influence of social network users is based on the principle of user interactive behavior similarity with the study of the PageRank algorithm, specific basis is as follows:

(1) Any micro-blog users have fans and attention, pay attention to other people similar to chain in web pages, fans similar to chain out web pages. The web pages structure of the network is similar to the micro-blog network.

(2) Evaluation of the influence of the user in the micro-blog, in fact is to estimate the user's ranking in the micro-blog network. PageRank is a classic Webpage ranking algorithm, and according to the Google Webpage search algorithm based on PageRank algorithm has been very mature and achieved great success.

PageRank algorithm in the calculation of the web page ranking and estimate the influence of micro blog users are different. The micro-blog relationship network and web pages have a similar topology, but the application environment is very different. In micro-blog user relationship, micro-blog user's influence not only with the number of fans, as well as user forwarding frequency and activity degree and so on. If the direct use of PageRank algorithm to estimate the influence of micro-blog users can not reflect the real situation. So we can improve the PageRank algorithm to adapt to the micro-blog environment by adding activity degree and forwarding rate.

Improved PageRank algorithm

Activity degree

A user's activity is determined by the number of users and the ability to spread the user's micro-blog and the number of users released microblogging and comments and the number of forwarding microbloging. The user's number of fans on behalf of the number of user's edges, the more fans on behalf of the large range of information dissemination, the more stronger of the user's transmission capacity. The transmission ability of the user's microblog is determined by the number of times of the user's micro-blog forwarding and comments. The spread of micro-blog information is showing a cascade of propagation, the average propagation depth of a micro-blog is about 3.5[4]. We will analyze the user activity degree evaluation model below.

The user u is assigned to the PageRank value of the user V, according to the user u to the active degree of user V to allocate. User v for the active degree of user u is:

$$A_{(u,v)} = \frac{A_u}{\sum_{i=1}^{n} A_i} \quad (2)$$

The A_u represents activity degree of user u, $\sum_{i=1}^{n} A_i$ represents the activity degree of the i fans of the user u. The n represents the number of fans of user u. The activity degree of a single user can be expressed as:

$$A_u = \sum \omega_{(i,j)} \cdot A_{(i,j)} + C \quad (3)$$

The $\omega_{(i,j)}$ in formula (3) represents the weight of each index. The $A_{(i,j)}$ represents the j impact index of the user. The C represents spread layer[5] series of user micro-blog. The weight $\omega_{(i,j)}$ can be made by number of fans, the user can be user communication ability, micro-blog users, comments, forwarding number micro-blog micro-blog number obtained by AHP. The frequency of the user's release can be calculated by the total number of micro blogging released by the first and the last of the time, and the number of user's forwarding and comments and the number of fans can

be directly obtained in the data collected. Evaluation of active users take full account of the user of the number of fans, microblogging release frequency and other factors, and these factors reallocation in accordance with certain rights. If the user has not logged micro-blog long time to publish active information, no comment forwarding, then the user in accordance with the level of consideration of the various indicators will fall. It will lead $A_{(i,j)}$ decline, and users do not have active information, the number of users of the spread of microblogging C will also decline. Zombie users can be excluded interference PageRank algorithm rankings.

Forwarding rate

In social networks, the user of information forwarding is the main way of information dissemination. Users forwarding messages will be see by attention node. By analogy, this message could be seen by all social network users. So we can see information forwarding is the main way of information dissemination. The research on the forwarding rate among the users is very significant for the research of social network user's influence[6].

The degree of user's attention determines the user's forwarding rate, by calculating the average value of the forwarding rate of all fans, the forwarding rate of $d(t)$ can be obtained:

$$d(t) = \frac{\sum_{(u,v)\in E} W_{(u,v)}(t)}{N_E} \tag{4}$$

The N_E in formula (4) represents the friends number of Edata set at time t. $W_{(u,v)}(t)$ represents (u,v) relationship attention degree at time t. The $W_{(u,v)}(t)$ formula is:

$$W_{(u,v)}(t) = \frac{N_{(u,v)}(t)}{T_{(u,v)}} \tag{5}$$

The $N_{(u,v)}(t)$ in formula (5) represents at t time period user u transmit message times of user, The $T_{(u,v)}$ represents the total number of messages for the user u to forward the user v. In the PageRank algorithm, the d represents the damping coefficient, and the forward rate is used to represent the damping coefficient.

Improved algorithm based on PageRank algorithm

The basic idea of PageRank algorithm is to distribute the PageRank value of the web page to the chain out page. So the PageRank algorithm model is applied to the micro-blog user influence model in the time has obvious flaws. In this paper, the forwarding rate is used as the weight of PageRank algorithm, and the activity degree of the user as the PageRank value of the index evaluation for each user. The two indexes are added to the PageRank algorithm to improve and optimize the algorithm. The improved PageRank algorithm as follows:

$$PR(u) = \big(1 - d(t)\big) + d(t) \times \sum_{v\in f(u)} A_{(u,v)} \times PR(v) \tag{6}$$

The formula (6) is the improved PageRank algorithm. This improved algorithm is based on traditional PageRank algorithm. The improved algorithm valuate the user's influence considering the activity degree and forwarding rate. The improved PageRank algorithm not only takes into account attention degree of users and the number of fans these static index, and also takes into account the dynamic behavior of the micro-blog user's forwarding and comments. Next, this article will be compared according to the experiment verify the validity of the improved algorithm.

Experiment results and analyses

Select a certain time t for calculation, compare in t time improved PageRank algorithm of rankings ranking with the original PageRank algorithm. We select the top 10 users to compare, as shown in Table 1.

Number	User name	Improved PageRank	PageRank
1	Happy Zhang jiang	1	4
2	Ye Liu	2	1
3	Nanpai Sanshu	3	95
4	Chen Li	4	30
5	Chao Deng	5	6
6	Kai Zheng	6	93
7	Mix teacher Lin Xin	7	14
8	Xing Jin	8	86
9	Na Xie	9	2
10	Yang Yang icon	10	7

Table 1

The same time improved algorithm and the original algorithm results comparison

It can be seen that the improved algorithm and the original algorithm in the rankings are very different. This is due to the active degree and forwarding rate as the evaluation index, not just look at the relationship. In ranking the gap larger Nanpai Sanshu user analysis, we found that at time t has a large number of users pay attention to Nanpai Sanshu and fans forwarding rate has significantly improved. Other rankings change is also similar to Nanpai Sanshu. The ranking change is obvious, as shown in Figures 1 shows the comparison between t1 and t6 time:

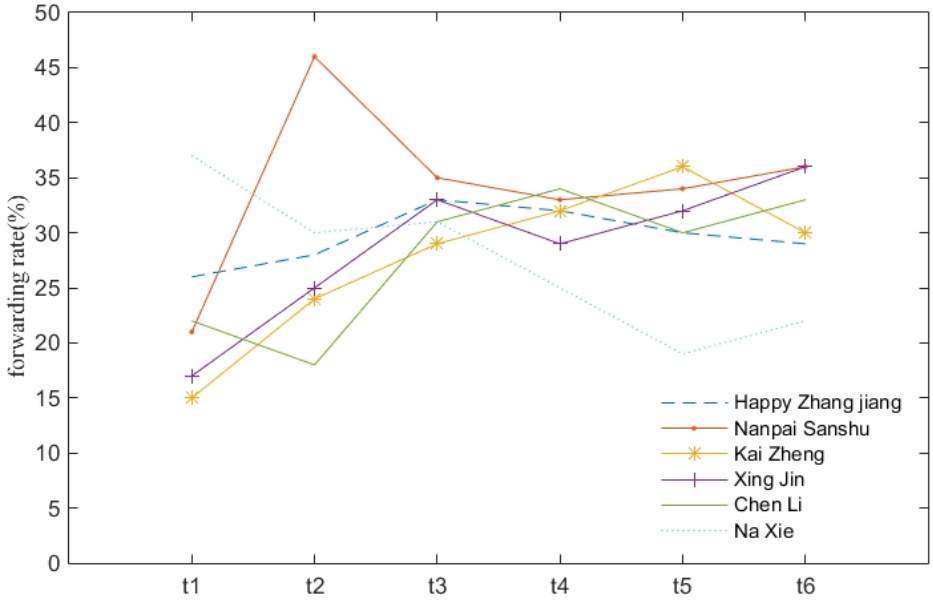

Fig.1. User forwarding rate between t1 and t6

This whole conform to the reality of life with the increased number of new fans, forwarding rate will increase. So the relationship between a certain period of time the number of fans and the growth of user influence is proportional.

Below I'll compare t1 with t2 use improved PageRank to ranking comparison. Time t1 and t2 between the 1 week time in accordance with the user influence ranking. Experimental results are shown in Table 2.

Number	User name	t2	t1
1	Chao Deng	1	1
2	Zhiqiang Ren	2	8
3	He Chen	3	5
4	angelababy	4	6
5	Na Xie	5	36
6	Bingbing Fan	6	9
7	Jiong He	7	7
8	Chen Li	8	3
9	Mi Yang	9	13
10	Tong Liu	10	21

Table 2

The ranking results use improved PageRank at different time

It can be found in Table 2, the change of the user's influence at different time is obvious. In order to analyze the causes of the changes in different times, ranked in the top 10 users to analyze than to find that the number of new fans are not very obvious but these changes are due to the user's activity has changed. The ranking of Na Xie is due to increased number of active information. The following figure illustrates a chart ranking increased significantly active user information to make the statistics, we find that according to the forwarding rate calculation formula, the more active user information, the information will increase the number of fans forwarded, so the user's forwarding rate increased, the user's ranking rose. The vertical coordinates represent the number of active information released every day, and the horizontal coordinates represent the time interval of a week, as shown in Figures 2:

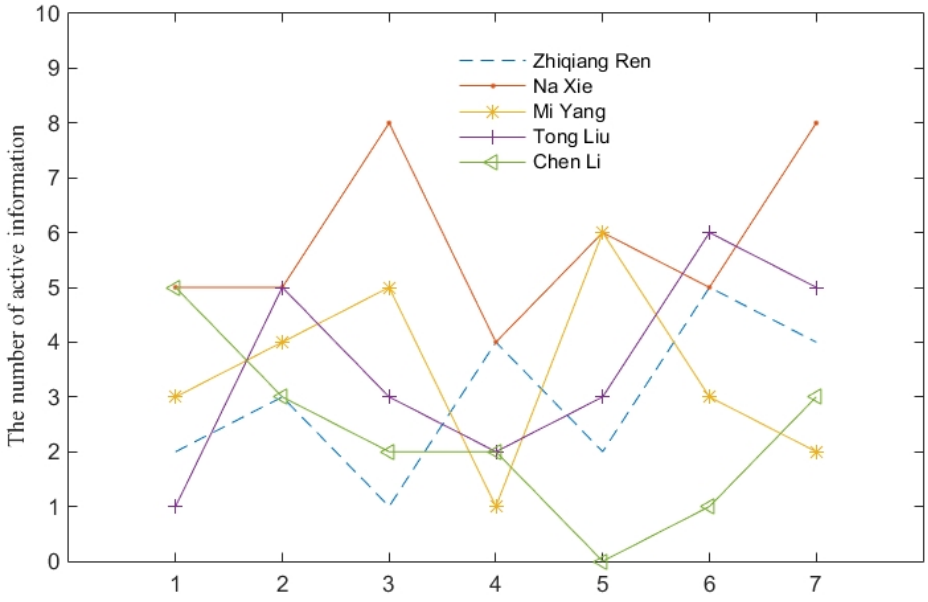

Fig.2. Active information statistics

After comparing PageRank algorithm with the improved PageRank algorithm at time t, the improved algorithm has a big difference with the original PageRank algorithm. The main reason is that the impact of user's forwarding rate and user activity is considered in the original PageRank algorithm. When analyzing different times improved PageRank algorithm ranking, influence by the user forwarding rate and activity degree. Ranking occurred corresponding changes, but mention is the improved algorithm in real time. Analysis from these two aspects, reflects the real-time and accuracy of the improved algorithm.

Conclusions

User influence as a hot research field of social network research. So through the discovery of social networks have a high influence of user community, can provide a theoretical basis for public opinion management, business promotion, has a very high practical significance. Based on the analysis of the influence of micro-blog users, this paper proposes an improved PageRank algorithm to study the influence of social network users. The algorithm is more in line with the characteristics of social network users, the influence of micro-blog users consider not only its dynamic characteristics, but also consider the static characteristics, make the algorithm more reasonable. The algorithm considers the user's activity degree, but also considers the forwarding rate of fans, from the fans forwarding rate as well as the user activity degree to comprehensively measure the influence of the user. After comparing and analyzing the rank of the micro-blog users, it is found that the algorithm is more powerful compared with the original PageRank algorithm.

References

[1] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, Bernardo A. Huberman. Influence and Passivity in Social Media. CoRR, 2014, Vol.abs/1008.1

[2]Heidemann J, Klier M, Probst F. Identifying key users in online social networks:A PageRank based approach. ICIS, 2010:79

[3]Haoran Wu, Gongwen Xu. Study on the influence of the cooperation network based on the PageRank algorithm. ICEECE，2015:37

[4] LI Zhiying, YANG Wu, XIE Zhijun. Summary of PageRank Algorithm. Computer Science, 2011, 38(10A).

[5]Jing Li, Jinli Guo. Research on the influence of micro blog users based on improved PageRank algorithm. China Newspaper Industry 2013.02:027.

[6] Ling Zhang, Zheng Qin. The Improved Pagerank in Web Crawler[C]. The 1st International Conference on Information Science and Engineering ICISE. Piscataway,N.J.:IEEE Press,2009:1889-1891.

[7] Yuan Fuyong, Feng Jing, Fu Qianqian. Micro-blog users influence index model. New Technology of Library and Information Service, 2012(6).

[8]Faloutsos M, Karagiannis T, Moon S. Online social networks[J]. IEEE NetWork，2010, 26(5):4-5.