To Find Key Node in Weibo with the Algorithm of UserRank Algorithm

Xuewei Hu [1, a], Xiyu Liu [2, b]

[1] College of Management Science and Engineering, Shandong Normal University, Jinan 250014, China;

[2] College of Management Science and Engineering, Shandong Normal University, Jinan 250014, China.

[a]huxuewei486@163.com, [b]sdxyliu@163.com

Abstract

Sina Weibo has now become one of the largest social network platform in China. Sina Weibo platform produces a large number of new user registrations and numerous new Weibo every day. The ways of summarizing and analysising of Sina Weibo contents and methods to find the key users in about three hundred million Weibo users are roughly divided into two categories: one is simply to determine the users' Weibo ranking by number of fans; the other is to use PageRank algorithm to sort. Because of the difference between the two sorting methods, this paper proposes an improved UserRank algorithm based on PageRank algorithm. UserRank algorithm is a way of evaluating the operations of link behavior between Weibo users,and then , we calculated the UserRank value of the user in specific time, according to the different values sort micro-blog users, so as to find the key nodes in this network structure.

Keywords

Sina weibo, UserRank, PageRank Algorithm, Key user identification.

1. Introduction

Online social network has been essential to everybody recent years. Sina Weibo as one of the biggest social web sites similar to Twitter is getting more and more attention.Weibo user registered on this web site can optionally look through blogs with web pages, WAP pages, mobile clients, and so on .Weibo is also a place where users share words, photos and even videos, what's more users could also interact with each other via following , forwarding ,commenting ,liking and chatting .

Contrast to some existing information platform, new social network platform such as Sina Weibo have unmatched advantage ,for example the Real time effect, emotion prediction function, openness, high information dissemination speed and so on .At present ,the researches of such social network platform are still in its infancy ,most of them concentrating on information dissemination and burst topic detection. The analysis of user importance is a basic study area as well as a research focus. Key users usually play an irreplaceable role in a social network platform. To identify the key users is crucial in understanding the process of information dissemination, analyzing user behavior and other research areas.

Because of associating with the way of following, forwarding, commenting, and liking, we can consider Sina Weibo as a data set with a mesh topology and data viewed as vertices of a graph ,the edges are relations mentioned above. Each vertices links to at least another one. Previous studies show that data set with mesh topology has high reliability and show relationships among vertices.To identify the key users is tough among such a complex network. According to graph theory, Sina Weibo can be abstracted to a graph , so identifying the key users is equal to look for key nodes in this diagram.

In this article, we consider that key users in online social networks are users who have the characteristics of attracting other users and disseminating information. The method of identifying key users must minimize the impact of extraneous factors and respond to user importance in a more rational way. So we use the dynamic local impact model to define the UserRank as the basis for ranking users ' importance, and key users can change their importance through the influence and manipulation of neighboring users in social networks, thereby changing their UserRank value, the higher the value, the greater the importance of users in the network. Sina Weibo is one of the largest online social networking platforms in China with a complete network-like user cluster, so we chose to experiment and apply the model on this platform.

2. Introduction to PageRank Algorithm

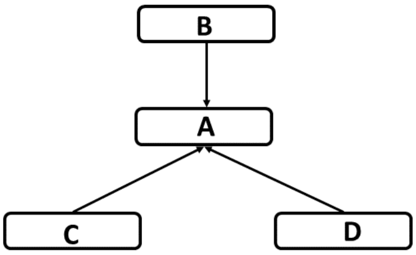
PageRank algorithm was proposed by Larry Page and Cheguer Brin in 1998, the algorithm is a very classic Web page ranking algorithm, the idea of this algorithm is based on network structure link analysis. PageRank determines the rank of a page through a hyperlink relationship between pages.

2.1 PageRank algorithm Principles

Google regards the link from A page to B page as the polls from A page to B page, depending on the source of the vote (or even source, that is, linking to A Page page) and the level of the voting target to determine the new level. In this way, PageRank evaluates the importance of the Web page based on the number of votes received on the B page , because voting on some important pages is considered to be of high value. So that the pages it links to are of high value. This is the core idea of the PageRank, and of course The actual situation of the PageRank algorithm is complex.

Suppose there is a collection of only 4 pages A, B, C and D ,. If all pages are linked toA, then A 's PageRank value will be the summation of PageRank value of B,C and D, as shown in the figure 1.

$$PR(A)= PR(B)+ PR(C)+ PR(D) \tag{1}$$



F ig.1

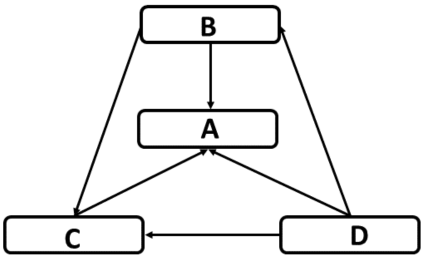


Fig. 2

Assume that B also links to C, and D links to the other 3 pages, shows in figure 1-2. A page cannot vote 2 times ,so B half-price each page. Similarly, only One-third of the D cast can be counted to the PageRank value of the " A ",the PR(A) is

$$PR(A)=PR(B)\times\frac{1}{2}+PR(C)\times 1+PR(D)\times\frac{1}{2} \tag{2}$$

The PR value of one page is the total number of links to the page . Use L (X) to indicate the number of linked pages that are linked to other pages of a page, and take Fig. 2 For example, to calculate the PR for each page in the diagram value is

$$PR(A)=PR(B)\times L(B)+PR(C)\times L(C)+PR(D)\times L(D) \tag{3}$$

$$PR(A)=PR(B)\times L(B)+PR(C)\times L(C)+PR(D)\times L(D) \tag{4}$$

$$PR(A)=PR(B)\times L(B)+PR(C)\times L(C)+PR(D)\times L(D) \tag{5}$$

$$PR(A)=PR(B)\times L(B)+PR(C)\times L(C)+PR(D)\times L(D) \tag{6}$$

2.2 Application of PageRan

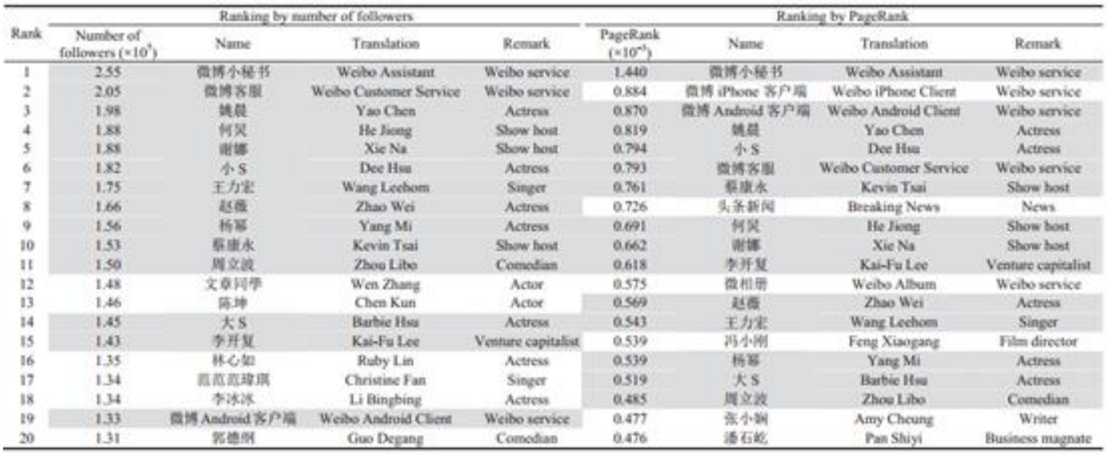
PageRank algorithm is well applied in many fields because of its superior computational performance. The main application fields are the order of importance of academic papers, the order of importance of the authors of academic papers, extraction of key words and sentences, etc.

2.3 Advantages and Disadvantages of PageRank

None of the algorithms are perfect, PageRank algorithm is no exception. The main advantage of the PageRank algorithm is that it is a static algorithm unrelated to the query, and that the PageRank value of all pages is obtained by off-line calculation, which greatly reduces the time of query response. The disadvantages are: (1) PageRank algorithm emphasizes the old page. Because old pages are more likely to be linked to other pages, the fact is that new pages may have better information prices. (2) PageRank the algorithm emphasizes the Web sites end with. com. Because this kind of website often is a comprehensive website, nature can obtain more link than other type of website, but the fact that some professional website is more authoritative to the elaboration of the problem.(3) PageRank algorithm cannot distinguish whether hyperlinks in a Web page are related to or unrelated to the theme of a Web page, that is, the similarity on the content of the Web page, which can easily lead to topic drift problems.

3. UserRank algorithm based on PageRank to find the critical users in Weibo

In Sina Weibo and Twitter, we calculated the PageRank values of top 100 users with most fans and sorted them according to both PageRank values and the number of fans respectively, and the Piercassen correlation coefficient between the two results is 0.82 and 0.94[1] .We can notice that on the Sina Weibo platform, the result of relevance of importance sorting just according to PageRank values or the number of followers is far less than Twitter. By survey results, on the Sina Weibo platform , we choose the top 20 users according to the number of fans and PageRank values respectively and the result shows in Fig. 3.Therefore, in the Sina Weibo platform to order the importance of users can not simply use the number of fans or the PageRank algorithm, therefore, this article is based on the PageRank Algorithm, combined with the number of Sina Weibo users and other variables proposed an improved user importance ranking algorithm, that is, the UserRank algorithm.



Rank	Ranking by number of followers				Ranking by PageRank			
	Number of followers ($\times10^7$)	Name	Translation	Remark	PageRank ($\times10^{-3}$)	Name	Translation	Remark
1	2.55	微博小秘书	Weibo Assistant	Weibo service	1.440	微博小秘书	Weibo Assistant	Weibo service
2	2.05	微博客服	Weibo Customer Service	Weibo service	0.884	微博 iPhone 客户端	Weibo iPhone Client	Weibo service
3	1.98	姚晨	Yao Chen	Actress	0.870	微博 Android 客户端	Weibo Android Client	Weibo service
4	1.88	何炅	He Jiong	Show host	0.819	姚晨	Yao Chen	Actress
5	1.88	谢娜	Xie Na	Show host	0.794	小 S	Dee Hsu	Actress
6	1.82	小 S	Dee Hsu	Actress	0.793	微博客服	Weibo Customer Service	Weibo service
7	1.75	王力宏	Wang Leehom	Singer	0.761	蔡康永	Kevin Tsai	Show host
8	1.66	赵薇	Zhao Wei	Actress	0.726	头条新闻	Breaking News	News
9	1.56	杨幂	Yang Mi	Actress	0.691	何炅	He Jiong	Show host
10	1.53	蔡康永	Kevin Tsai	Show host	0.662	谢娜	Xie Na	Show host
11	1.50	周立波	Zhou Libo	Comedian	0.618	李开复	Kai-Fu Lee	Venture capitalist
12	1.48	文章同学	Wen Zhang	Actor	0.575	微相册	Weibo Album	Weibo service
13	1.46	陈坤	Chen Kun	Actor	0.569	赵薇	Zhao Wei	Actress
14	1.45	大 S	Barbie Hsu	Actress	0.543	王力宏	Wang Leehom	Singer
15	1.43	李开复	Kai-Fu Lee	Venture capitalist	0.539	冯小刚	Feng Xiaogang	Film director
16	1.35	林心如	Ruby Lin	Actress	0.539	杨幂	Yang Mi	Actress
17	1.34	范范范玮琪	Christine Fan	Singer	0.519	大 S	Barbie Hsu	Actress
18	1.34	李冰冰	Li Bingbing	Actress	0.485	周立波	Zhou Libo	Comedian
19	1.33	微博 Android 客户端	Weibo Android Client	Weibo service	0.477	张小娴	Amy Cheung	Writer
20	1.31	郭德纲	Guo Degang	Comedian	0.476	潘石屹	Pan Shiyi	Business magnate

Fig. 3 the top 20 users inWeibo

3.1 Application of PageRan

Because the sheer number of fans or PageRank algorithm fails to comprehensively analyze the importance of Sina Weibo platform's users, and we propose a more comprehensive measure of user importance named UserRank . This algorithm bases not only on the static index of user's fan number, but also behavioral indicators including forwarding, comment, point-praise, attention, collection. Weibo identification and Weibo number are also concluded.

3.1.1 Application of PageRan

Fig.4 shows a screenshot about a piece of Weibo it can be seen that a Weibo account has concerns, number of fans, number of Weibo, Weibo certification and other static indicators, as well as collection, forwarding, comments, point praise, attention, cancellation of attention and other behavioral indicators. This article selects all of the above metrics as a composition of the formula to calculate the UserRank value.

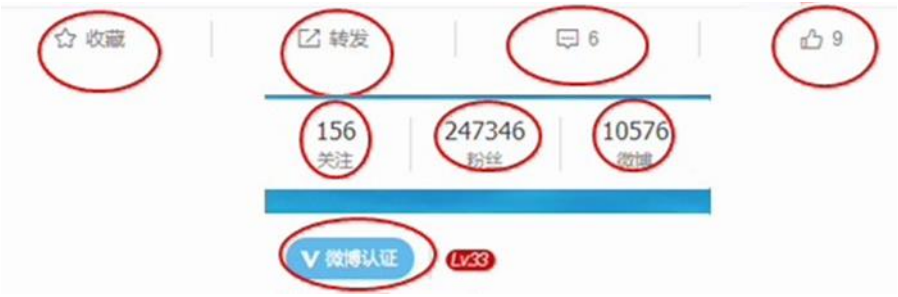


Fig. 4 a piece of Weibo

For numeric indicators (attention, number of fans, micro-bo number), we take its accurate values and for non-numeric indicators (micro-blog authentication, collection, forwarding, comments, point praise, attention, cancellation of attention, etc.), we carry out the dimensional treatment, according to the importance of indicators to give them the corresponding weights.

For the indicator of Weibo identification, it can be subdivided into government identification, enterprise identification, corporate organizations, media identification, celebrity identification, personal identification, and other types. Different certified users have different rights, such as institutional groups and media identification has the right to speak, enterprise identification and celebrity identification has a priority of search rights [2], therefore, according to the user identification rights, we give different identification users their weight values. The weights are calculated using the analytic hierarchy process with yaahp software, and the consistency test results are shown in the Table 1.

Table 1 consistency test results

Identification	Weights
enterprise identification	0.3560
government identification	0.2532
media identification	0.1596
celebrity identification	0.1407
corporate organizations	0.0068
personal identification	0.0236

In this way we get the respective weights of the six Weibo identification types. Using the analytic hierarchy process with yaahp software, we get the importance weights of the five behavioral indicators as shown in Table2 .

Table 2 weights of behavioral indicators

Behavioral indicators	Weights
Forwarded	0.4174
Comments	0.2634
Click to praise	0.1602
Follow	0.0975
Favorites	0.0615

3.1.2 Formula of UserRank

The model has time changes, an initial time is selected which is 0 , 1 is a time unit in which each user can perform only one operation (one of the 5 basic behaviors). So in the initial time, the importance I_i^0 of the each user i is

$$I_i^0 = \frac{F_{1i}^0}{F_{1i}^0+F_{2i}^0} \cdot \frac{F_{1i}^0}{N^0} + \frac{WB_i{}^0}{Avg^0} + a_i \qquad (7)$$

In the formula, F_{1i}^0 means the number of Weibo users who followed user i in time 0, F_{2i}^0 is the number of Weibo users followed by user i in time 0, N^0 is the total number of Weibo user in time 0, $WB_i{}^0$ is the number of Weibo processed by user i in time 0, Avg^0 is the average number of Weibo processed by every user in time 0, $a_i$is weights of identification of users. $\frac{F_{1i}^0}{F_{1i}^0+F_{2i}^0}\cdot\frac{F_{1i}^0}{N^0}$ shows the popularity of user i in time 0 and $\frac{WB_i{}^0}{Avg^0}$ means the activity of user i in time 0.

In time 0 ,the UserRank of user i is

$$\mathrm{UserRank(i)}^0 = \frac{I_i^0-\mathrm{M}in(I_b^0)}{\mathrm{Max}(I_a^0)-\mathrm{M}in(I_b^0)} \tag{8}$$

Max (I_a^0) and Min (I_b^0) $(\mathrm{a} \neq \mathrm{b})$ are the biggest and smallest importance among all users. $\mathrm{UserRank(i)}^0$ can short for $\mathrm{UR(i)}^0$.Iterate user importance from the initial state, on time t+1, I_i^{t+1} is

$$I_i^1=\sum_{k=1}^{N^0}(ZF_{\mathrm{ik}}^1+PL_{\mathrm{ik}}^1+DZ_{\mathrm{ik}}^1+GZ_{\mathrm{ik}}^1+SC_{\mathrm{ik}}^1)+UR_i^0\ (\mathrm{t=0}) \tag{9}$$

$$I_i^{t+1}=\sum_{k=1}^{N^t}(ZF_{\mathrm{ik}}^{\mathrm{t}}+PL_{\mathrm{ik}}^{\mathrm{t}}+DZ_{\mathrm{ik}}^{\mathrm{t}}+GZ_{\mathrm{ik}}^{\mathrm{t}}+SC_{\mathrm{ik}}^{\mathrm{t}})+UR_i^t\ (\mathrm{t>0}) \tag{10}$$

$\mathrm{ZF}_{\mathrm{ij}}^{\mathrm{t+1}}$is the importance of user i forwards a Weibo in time t+1, $\mathrm{PL}_{\mathrm{ij}}^{\mathrm{t+1}}$ is the importance of user i comments a Weibo in time t+1 , DZ_{ij}^{t+1} is the importance of user i clicks to praise a Weibo in time t+1, GZ_{ij}^{t+1} is the importance of user i follows a Weibo in time t+1,SC_{ij}^{t+1} is the importance of user i Favorites a Weibo in time t+1.

$$ZF_{\mathrm{ij}}^{t+1}=\begin{cases} UR_i^t\cdot w_{ZF}\cdot r_{ij} & r_{ij}=1 \\ 0 & r_{ij}=0 \end{cases} \tag{11}$$

$$PL_{\mathrm{ij}}^{t+1}=\begin{cases} UR_i^t\cdot w_{PL}\cdot \eta_{ij} & \eta_{ij}=1 \\ 0 & \eta_{ij}=0 \end{cases} \tag{12}$$

$$DZ_{\mathrm{ij}}^{t+1}=\begin{cases} UR_i^t\cdot w_{DZ}\cdot \sigma_{ij} & \sigma_{ij}=1 \\ 0 & \sigma_{ij}=0 \end{cases} \tag{13}$$

$$GZ_{\mathrm{ij}}^{t+1}=\begin{cases} UR_i^t\cdot w_{GZ}\cdot \varphi_{ij} & \varphi_{ij}=1 \\ 0 & \varphi_{ij}=0 \\ -UR_i^t\cdot w_{GZ}\cdot \varphi_{ij} & \varphi_{ij}=-1 \end{cases} \tag{14}$$

$$SC_{\mathrm{ij}}^{t+1}=\begin{cases} UR_i^t\cdot w_{SC}\cdot \xi_{ij} & \xi_{ij}=1 \\ 0 & \xi_{ij}=0 \end{cases} \tag{15}$$

w_{ZF}, w_{PL},w_{DZ},w_{GZ},w_{SC} are weights of different behaviors and they have been calculated above. After calculating the user importance, normalize I_i^{t+1} to control its value in [0,1]. And the value we get is the UserRank in time t+1, marked as UR_i^{t+1} and

$$UR_i^1=\frac{I_i^1-\mathrm{M}in(I_b^1)}{\mathrm{Max}(I_a^1)-\mathrm{M}in(I_b^1)}\ (\mathrm{t=0}) \tag{16}$$

$$UR_i^{t+1}=\frac{I_j^{t+1}-\mathrm{M}in(I_b^{t+1})}{\mathrm{Max}(I_a^{t+1})-\mathrm{M}in(I_b^{t+1})}\ (\mathrm{t>0}) \tag{17}$$

After the above steps we can get Sina Weibo users' UR_i^{t+1} value in a specific moment, sorted again, Weibo users are ranked, key users can be clearly found in this network-like data structure.

3.2 An instance of UserRank Algorithm

To better illustrate the UserRank algorithm, this section extracts a small number of examples from the Sina Weibo platform to compute their respective UR values. Extracted users and their number of static metrics (up to 2018 year March) as shown in Table 3 and the behavior relationships that occur within a time unit are shown in the Fig. 5 .

Table 3 Extracted users and their number of static metrics

Nickname	F_{2i}^{0}	F_{1i}^{0}	${WB_i}^0$	identification
A	203	1980475	15653	corporate organizations
B	545	2430895	20515	media identification
C	1087	2428	10900	personal identification
D	564	1860	13247	personal identification
E	723	477718	3184	corporate organizations
F	349	41	1000	personal identification
G	125	5864932	8119	enterprise identification
H	112	232404	505	enterprise identification

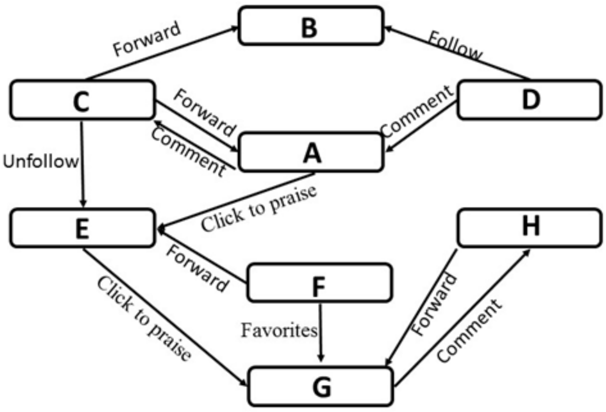


Fig.5 behavior relationships

Table 3 shows the initial static values and behaviors in Fig. 5 happened in one time unit. The UserRank values in time 0 are showed in Table 4.

Table 4 UserRank values in time 0

Nickname	number	$\mathrm{UR(i)}^0$	rank
A	1	1	1
B	2	0.4752	4
C	3	0.7252	2
D	4	0.5879	3
E	5	0.1245	6
F	6	0	8
G	7	0.1224	6
H	8	0.3693	5

After operations in Fig.5, the new UserRank values in time 1 are showed in Table 5.

Table 4 UserRank values in time 0

Nickname	number	$\mathrm{UR(i)}^0$	Original Ranking	$\mathrm{UR(i)}^1$	rank
B	1	1	1	1	1
C	2	0.4752	4	0.5306	3
A	3	0.7252	2	0.8588	2
D	4	0.5879	3	0.4682	4
E	5	0.1245	6	0.2264	6
F	6	0	8	0	8
G	7	0.1224	6	0.1750	7
H	8	0.3693	5	0.3507	5

As can be seen from the table, after an iteration, the user's UR the value has changed and the sequence of users corresponding to UR values has changed a lot. This allows key users in this network-like structure data to be selected based on the user's latest UR rankings.

4. Conclusion

Based on the above methods, the complex of user relationships in social networking sites is abstracted into weighted forward graphs, which can calculate the importance of the users by calculating the correlation of graph, and then finds the key users. However, the method presented in this paper still has some shortcomings and can be improved.

The problem is that the importance of users and UR values are sorted two times in one time unit respectively, which undoubtedly increases the computational difficulty. Secondly, in the Sina Weibo platform, some registered accounts do not operate after the attention of other accounts in a considerable period of time (commonly known as "zombie powder"), the existence of such accounts also affect the accuracy of the UR value calculation.

The formula proposed in the third chapter uses both static and dynamic variables, but as the social networking platform continues to escalate, the real social networking platform continues to produce new features. For example, a user can mention (@) Other people at the time of the creating Weibo,or add a site links (link to internal or external of this social platforms), or make headlines by other means. These new features will, to some extent, affect the importance of the user, so the formula in chapter three can further consider these new factors and continually improve The calculation of the UR value.

Finding key users in complex network-like structure data is a difficult problem. In this paper, a model based on time changing and users’ static and dynamic index is proposed to identify key users in online social network platform. After analyzing the data of Sina Weibo, we find that there are up to five associations among users, so we assign the action behavior of these five kinds of users to the importance of them, so that the behavioral index becomes a convenient numerical index. On this basis, we turn the purpose of identifying the key users in the social networking platform to calculate the UR value for each user. This is both an inheritance and an improvement to the classic PageRank algorithm. The instance results of the 3.2 section show that this model, which is based on time change and user static and dynamic metrics, can describe the relationship between users and compute the importance of users in a social networking platform in a more efficient way.

Acknowledgements

Natural Science Foundation.

References

[1] Wentao Han, Xiaowei Zhu, Ziyan Zhu. A Comparative Analysis on Weibo and Twitter, TSINGHUA SCIENCE AND TECHNOLOGY, vol.2(2016).2, 1-16.

[2] Sergey Brin , Larry Page, The PageRank Citation Ranking : Bringing Order to the Web, Google search engine. http://google.stanford. Edu.

[3] B. Viswanath , A. Mislove , M. Cha, and K. P. Gummadi , On the evolution of user interaction in facebook , in Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN’09), vol. 39(2009),37-42.

[4] WANG Nan, SUN Qindong, ZHOU Yadong , et al,A Study on Influential User Identification in Online Social Networks,Chinese Journal of Electronics , Vol.25(2016), 467-473.

[5] Wang Zhiwei, Wu Shuxia. The research and improvement of PageRank algorithm application in the sort of document retrieval , Information Systems , vol. 39(2016),124-144.

[6] Huang Decai, Qi Huanchun . PageRank Algorithm Research [J]. Computer Engineering, vol. 32(2006), 145-146.

[7] G.Z. Dong , R.G. LI and W. Yang, “Microblog burst keywords detection based on social trust and dynamics model”, Chinese Journal of Electronics, Vol.23(2014),695–700.