Summer 5-27-2016

# A PAGERANK-BASED MINING ALGORITHM FOR USER INFLUENCES ON MICRO-BLOGS

Guo-Jun Mao
*Central University of Finance and Economics*, maoguojun@cufe.edu.cn

Jie Zhang
*Central University of Finance and Economics*, 1911364383@qq.com

# A PAGERANK-BASED MINING ALGORITHM FOR USER INFLUENCES ON MICRO-BLOGS

Guo-Jun Mao, Information Scholl, Central University of Finance and Economics, Beijing, P R China, maoguojun@cufe.edu.cn

Jie Zhang, Information Scholl, Central University of Finance and Economics, Beijing, P R China, 1911364383@qq.com

## Abstract

*With the development of Web technology, the Micro-Blog has become one of the most popular social platforms, and calculating and ranking the influences of the users on Micro-Blogs has been issuing an important research problem. Through improving the traditional the PageRank model, this paper presents a called PR4MB (PageRank for Micro-Blog) algorithm, which can obviously improve mining precisions for evaluating user influences on a Micro-Blog. While considering user link relations like the PageRank method, the PR4MB algorithm also takes attention to the activity, quality and credibility of a user on a Micro-Blog, so it constructs a dynamic mining model for user influences on a Micro-Blog by evaluation user online behaviors. The experimental results show that PR4MB algorithm, in comparing with the traditional PageRank algorithm, can more truly reflects the actual influences of different users on a Micro-Blog.*

*Keywords: PageRank, User influence, Micro-Blog, Twitter, Ming algorithm.*

# 1 INTRODUCTION

With the advent of the era of big data and the application of Web 2.0 technology, online social network has been developed rapidly. As a kind of important social platforms, the Micro-Blog (Wei-Bo) has become an important carrier of public opinion in different aspects, such as society, politics , lives and so on. Twitter, where is one of the most popular blogging platforms in the word. According to statistics, up to July 2014, Twitter's active registered users have exceeded 600 million, and the daily average number of new users is 135000. However, the influences of the blog users on a Micro-Blog are very different, so evaluating and ranking the blog users is an important issue. The influence of a user on a Micro-Blog is related to many factors, including not only link relations to other users but also its online behaviours such as its activity, credibility and the quantity and quality of its blog posts.

The PageRank mode (Page et al, 1999) was proposed by Larry Page and Sergey Brin, the two founders of Google. The original cause of the model is to achieve the web page ranking, which has been an important ingredient in the ranking of search results used by Google. However, it has been explored for analyzing social networks in recent years (Tunkelang et al, 2009). In fact, the PageRank model is essentially a computing technology for directed graph's node level, so it is nature to apply to the influence of users on Micro-Blogs.

In 2009, Tunkelang et al. constructed a link-based directed graph and used the PageRank model to achieve the Twitter user's influence ranking (Tunkelang et al, 2009). In 2010, Weng et al also successfully created a good-friend relationship graph from Twitter, and proposed an algorithm called TwitterRank by making use of the PageRank model (Weng et al, 2010). In 2011, Weng et al improved its TwitterRank algorithm that added user interests to the evaluation system, which can better realize the analysis of user influence in a specific topic discussion (Weng et al, 2011).

Recently, there are three main topics have been greatly concerned in mining user influences on Micro-Blogs:

(1) The analysis of interaction behaviors in Micro-Blog users, where social network always takes a special focus (Yin et al, 2014; Kempe et al, 2003).
(2) Using the theory of information dissemination to study the dynamic behaviors of the network in the Micro-Blog in order to find the key users (Xiong, 2014).
(3) Conducting the data analysis models from the view of data mining, and find the implicit interaction rules or user behaviors in the Micro-Blog (Deng et al, 2014; Noordhuis, 2010).

These studies have a core technical support, that is, directed graph analysis. Generally, the users in a Micro-Blog are always as the nodes of a directed graph, and their social relations or the spreads of blog posts can been converted into the arcs of the directed graph. Therefore, in this sense, the PageRank model can be used to calculate user influences of the Micro-Blog.

Based on the PageRank model, the technical solution of this paper is simply described as follows: First, using the social network concepts and its analysis tools, an directed graph , the logical analysis space of a Micro-Blog, is built; Then, the influence factors of the user behaviors caused by the blog communication are analyzed, a integrative evaluation model of the user weightiness is constructed; Finally, through improving the process of the traditional PageRank algorithm, a dynamic mining algorithm for calculating the user influences on the Micro-Blog is designed.

# 2 PRELIMINARIES TO THE PAGERANK MODEL

As is known, the original PageRank model is proposed to evaluate the importance of web pages in a web site. Simply said, the PageRank model employs such an idea that a web page is more important if more other web pages link to it. Further, within the PageRank concept, the rank of a page is related to

all pages that link to it, so the rank of a page is always determined recursively by the ranks of those pages that link to it.

**Definition 1 (Rank of Page).** Given a directed graph $G=<V, A>$ related to a Web site. The rank of a page $U$ ($\in G$) can be calculated as follows:

$$PR(U) = (1-d) + d \sum_{v \in I(U)} \frac{PR(V)}{O(V)}, \qquad (1)$$

Where:

- $PR(U)$ is the page $U$'s rank value;
- $I(U)$ is the set of these pages that link to $U$, i.e. any $V \in I(U)$, $<V,U> \in A$;
- $O(V)$ is the number of outbound links from page $V$;
- $d$ is a damping factor which can be set between 0 and 1.

According to Definition 1, the rank of a page $V$ impacts on its inbound page $U$ by weighting rank that divides its original rank $PR(V)$ by the number of outbound links $O(V)$ of page $V$. This means that the more outbound links a page $V$ has, the less will page $U$ benefit from page $V$. Thereby, the rank of a page can be inferred from the weighted ranks that all link to it. Node that Formula (1) is called Random Surfer Model (Rogers, 2002), as it uses damping factor $d$. Such, even if a page has no inbound page, it can still be jumped into with probability (1-$d$), which can be effectively avoid stopping calculating on the way.

In fact, Formula (1) is a foundational description for calculating the page ranks in a Web site, there have been some computable algorithms to appear, where one of the most popular ways is based on the Transition Probability Matrix that is called the Google Matrix such that it is used in the Google search engine (Lin, 2009).

**Definition 2 (Google Matrix).** Given a directed graph $G=<V, A>$, its Google Matrix $M= (m_{i,j})$, where $m_{i,j}$ is the probability of page $j$ jump to the page $i$. That is, for $i, j \in G$, the Google Matrix related to the directed graph $G$ can be set as follows:

$$m_{i,j} = \begin{cases} 1/O(j), & when < j,i > \in A \\ 0. & otherwise \end{cases}, \qquad (2)$$

Where $O(j)$ is the number of outbound pages of $j$.

For example, for the directed graph shown in Figure 1, its Google Matrix can be obtained by Formula (2). Figure 2 gives such a Google matrix $M$.

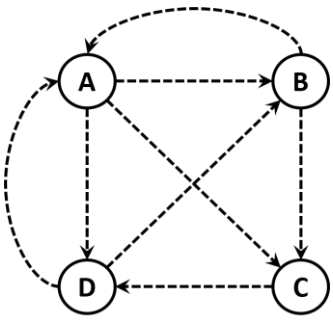


*Figure 1. A directed graph for page links*

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \end{bmatrix}$$

*Figure 2. The Google Matrix of Figure 1*

As a matter of fact, the ranks of all Web pages in Google engine are iteratively updating until they are stable. Therefore, calculating Rank Vectors is the main task of the PageRank model.

**Definition 3 (Rank Vector).** Given a directed graph $G$ and its Google Matrix $M$, and set $R= (r_1, r_2,... , r_n)^T$ is the vector that comprises of all ranks of the pages in $G$, where $n$ is the number of nodes in $G$. Then, the rank vector can be iteratively updated step by step until it is stable to a suitable value, by the following Formula (3) .

$$R_{i+1} = (1-d)*(1,1,..,1)^T + d*M*R_i, \qquad (3)$$

# 3 PR4MB：PAGERANK FOR MICRO-BLOGS

The traditional PageRank model can help evaluating the influences of Micro-Blog users, but its effectiveness has been questioned (Tunkelan, 2009; Weng, 2010). For example, the traditional PageRank model only considers link relations, and it is actually supposed the same importance of all users. However, in a Micro-Blog, users are playing different roles. Therefore, we should take more attentions to user characters such as user activities and confidences.

All of first, the user activity can be considered as one of the important factors to the user influences. Generally, the more blogs a user publishes, the more activity this user has in the Micro-Blog, so a typical measure for calculating a user active degree can be described as Definition 4.

**Definition 4 (User Activity).** Given a user $i$ in the investigated Micro-Blog, its activity can be calculated as follows:

$$Activity(i) = n_i / N, \qquad (4)$$

Where $n_i$ is the number of blogs that user $i$ publishes, and $N$ is the number of all users write on this Micro-Blog.

In addition, the user activity only reflects the blog quantities of a user but ignoring its blog qualities, so it should be necessary to evaluate the blog quality of a user in order to enhance its effectiveness (Yin, 2011). In general, the qualities of the blogs are directly related to the subject of the observation. For example, if we observe the social problems or issues, then the users who pay greater attentions to these topics should be evaluated into a higher level. On the contrary, these users whose interests in other topics such as commodity sales should be given in a lower level. Following this idea, the following Definition 5 presents a simple evaluating method of the user qualities on a Micro-Blog.

**Definition 5 (User Quality).** Given a user $i$ and the concerned topic $T$ in the investigated Micro-Blog, its quality can be calculated as follows:

$$Quality(i) = m_i / M_i, \qquad (5)$$

Where $m_i$ is the number of blogs related to $T$ that User $i$ publishes, and $M_i$ is the number of all blogs that User $i$ write in the Micro-Blog.

There is another factor can also reflect a user influence to a great extent, which is its credibility. For example, people are used to believe what the authorities or truepennies say, and it is also so in theMicro-Blog word. In fact, this is a difficult job because the authority of a user can be related to many factors. However, as a Micro-Blogging platform often provide the certification mechanism, so a simple and effective method is just making use of these certifications. Definition 6 gives a simple evaluating method of a user credibility on a Micro-Blog.

**Definition 6 (User Credibility).** Given a user *i* in the investigated Micro-Blog, its credibility can be calculated as follows:

$$Credibility(i) = \begin{cases} 1/O(i), when \quad user \quad i \quad is \quad centified \\ 0, \qquad\qquad\quad otherwise \end{cases}, \qquad (6)$$

Where $O(i)$ is the number of outbound users of User *i*.

In fact, there have been some discussions in adding the other criteria beyond the link structure of users on the Micro-Blogs (Java et al, 2006; Cha et al, 2010), which can enhance the effectiveness of evaluating the user ranks. Therefore, we will modify the user weightiness before running the PageRank algorithm.

**Definition 7 (User Weightiness).** Given a user *i* in the investigated Micro-Blog, its weightiness can be calculated as follows:

$$w(i) = Activity(i) + Quality(i) + Credibility(i), \quad \ldots..(7)$$

Where $Activity(i)$, $Quality(i)$ and $Credibility(i)$ respectively represent activity, quality and credibility of User *i* that have been calculated by Definition 4 to 6.

Now, we can discuss the problem of improving the original PageRank model in order to mine the user influences on a Micro-Blog. A basic idea is just modifying the PageRank model by using the user weightiness settings.

**Definition 8 (Weighted Google Matrix).** Given a directed graph $G=<V, A>$, its Weighted Google Matrix $P=(p_{i,j})$ , where $p_{i,j}$ is calculated as follows:

$$p_{i,j} = \begin{cases} w(i)/O(j), & when < j,i > \in A \\ 0. & otherwise \end{cases} \qquad (8)$$

Where $O(j)$ is the number of outbound nodes of *j*.

Obviously, $p_{i,j}$ in Definition 8 is the value $m_{i,j}$ in Definition 2 multiples by the user weightiness $w(i)$, so a Weighted Google Matrix consider more factors than link structure of users, including of user activity, quality and credibility.Continuing to doing analyzing for Figure 1, if we have obtained the evaluation parameters of each node: $w(A)=1.2$, $w(B)=0.2$, $w(C)=0.8$ and $w(D)=0.1$, then the corresponding Weighted Google Matrix *P* can be obtained as shown in Figure 3.

$$P = \begin{bmatrix} 0 & 0.6 & 0 & 0.6 \\ 0.07 & 0 & 0 & 0.1 \\ 0.27 & 0.4 & 0 & 0 \\ 0.03 & 0 & 0.1 & 0 \end{bmatrix}$$

*Figure 3. The Weighted Google Matrix for Figure 1*

Using Weighted Google Matrix, the PageRank model can be modified, and become a novel mining model for Micro-Blog. Algorithm PR4MB in the following describes the main processing courses based on such an improving PageRank model.

**Algorithm PR4MB (PageRank for Micro-Blog)**

**Input:** Micro-blog user social network graph $G$; damping coefficient $d$; iteration termination condition $\mathcal{E}$.

**Output:** The influence vector of user node $R$.

**Process:**

1. Calculate the Google Matrix of $G$ : $M=(m_{i,j})$;
2. FOR i $\in$G DO
3. Calculate $Activity(i)$, $Quality(i)$, $Credibility(i)$ by Definition 4~5;
4. $w(i)= Activity(i)+Quality(i)+ Credibility(i)$;
5. ENDFOR
6. FOR $i \in G$ DO
7. FOR $j \in G$ DO
8. $p_{i,j} = w(i)* m_{i,j}$;
9. $P =( p_{i,j})$;
11. $R_0 = I$;
12. REPEAT
13. $R =(1-d) * I + d * P * R_0$;
14. $R_0 = R$;
15. Until $||R- R_0|| \leq \mathcal{E}$;
16. Return $R$ as the final influence vector.

In Algorithm PR4MB, Step 1 is to generate the Google Matrix $M$ like to do in the traditional Page Rank algorithm; Step 2~5 calculate user weightiness through evaluating the user's activity, quality and credibility; Step 6~9 generate the Weighted Google Matrix $P$; Step 10~15 iteratively produce the rank vectors, and it is sopped when two rank vectors is rather closed.

# 4 EXPERIMENT AND ANALYSIS

The experimental data comes from the Twitter, a popular social website. We use API got data from Twitter for nearly 5 years, then selected 3049 blog users to generate the social networks by the Gephi tool. Though 3049 users are only a small part of the total Twitter users, but the relationship of these users is relatively complete, so we can use them to test the effectiveness of our algorithm in this paper.

| No. | Blog Name | Certified | Blog No. | Follower No. | Friends No. |
|---|---|---|---|---|---|
| 1 | Oasis Feng | 0 | 6568 | 2293 | 86 |
| 2 | Stephen Colbert | 1 | 3532 | 7570195 | 1 |
| 3 | Bill Clinton | 1 | 278 | 3117936 | 10 |
| 4 | Fenng | 0 | 28778 | 81452 | 3485 |
| 5 | Google | 1 | 6179 | 10397244 | 433 |
| 6 | Bill Gates | 1 | 1578 | 20692667 | 166 |
| 7 | Kai-Fu Lee | 1 | 1197 | 1303670 | 142 |
| 8 | Chris Hadfield | 1 | 7499 | 1297021 | 66 |
| 9 | Barack Obama | 1 | 13235 | 56418972 | 644143 |
| 10 | NASA | 1 | 35883 | 9206040 | 231 |

*Table 1. The top ten users information table calculate by Traditional PageRank algorithm*

| NO. | Blog Name | Certified | Blog No. | Follower No. | Friends No. |
|---|---|---|---|---|---|
| 1 | Huffington Post | 1 | 393378 | 5425645 | 5567 |
| 2 | BBC News (UK) | 1 | 247665 | 3812666 | 104 |
| 3 | BBC News (World) | 1 | 195627 | 9221709 | 61 |
| 4 | The New York Times | 1 | 172389 | 15720420 | 984 |
| 5 | Washington Post | 1 | 131347 | 4154603 | 1164 |
| 6 | Wall Street Journal | 1 | 112394 | 5994789 | 1008 |
| 7 | The Associated Press | 1 | 102580 | 4883727 | 7339 |
| 8 | NASA | 1 | 35883 | 9206040 | 231 |
| 9 | NYT Opinion | 1 | 40129 | 149632 | 1401 |
| 10 | Chris Hadfield | 1 | 7499 | 1297021 | 66 |

*Table 2. The top ten users information table calculate by PR4MB algorithm*

Table 1 and table 2 respectively give the Top 10 users and their corresponding situations found by the traditional PageRank and the PR4MB algorithm, where the topic is about social issues for PR4MB.

As stated in Table 1 and 2, on the same data set, the mining results between the traditional PageRank algorithm and the PR4MB algorithm are very different. Though we can't still ensure to say which is better by them, some of the problems in Table 1 can be found by analysis as follows:

(1) In table 1, Oasis Feng is thought out the 1st important user, who is not only a non-certified user but also has 6568 blogs which is not much more than others. Through further tracking calculating process by the traditional PageRank algorithm, we found two main reasons that result in it: the first is that the followers of Oasis Feng have higher ranks; the second is that he has less links into others. Such, he gets more but sends less. However, its irrationality is obvious for such a Micro-Blog.

(2) In table 1, the ranked fourth is also a non-certified user and with a less followers than the ranked $5^{th}$ to $10^{th}$. Such a situation is questionable. The reason is that Fenng is famous IT reviewers, whose blog posts are mostly about IT technical discussions, which has greater authority in IT field. However he does pay more attentions to the social topics. Therefore, this exposes the defect that the traditional PageRank algorithm cannot well reflect the user differences and their biases.

In comparison, PR4MB algorithm consider the activity, quality and credibility of a user before interactively the user rank, which makes the user influence evaluating in an integrated and topic-oriented way. As shown in Table 2, the top 10 users are found out by PR4MB should be more closed to the real word on Twitter. They are all certified users as well as with enough posts, followers and friends. In addition, in order to assess the quality of the articles, we put the theme in social issues, so the PR4MB algorithm got the first ten users that are mostly the news media with official certifications which should have greater authoritativeness.

In addition, we also tested the execution time changes in the traditional PageRank and PR4MB algorithms with different numbers of users. The experimental data were obtained from the Twitter platform, which has 3049 users to form a social network. In order to track their time consumptions with different numbers of users in the traditional PageRank and PR4MB algorithms, we segment it into a few of different sub-graphs that respectively include 500, 1000, 1500, 2000 and 2500 users. The experiments was implement in a computer with quad-core CPU, 4G memory and Win 7 operating environment. Figure 4 gives the corresponding experimental results.

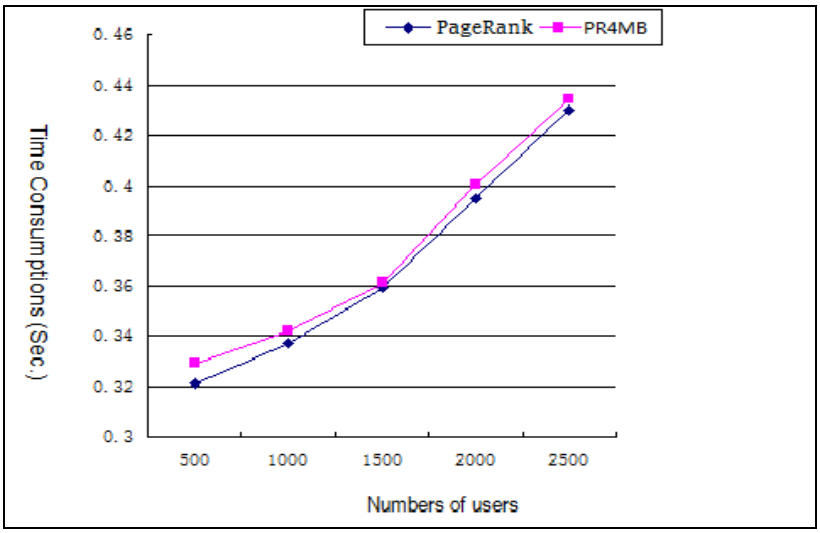


*Figure 4. Time comparison between PR4MB and PageRank algorithms*

Figure 4 shows that the execution time of the PR4MB algorithm is slightly higher than the PageRank algorithm doing in the same data capacity. This is because compared to the traditional PageRank algorithm, the PR4MB algorithm needs to calculate the user's weightiness values in advance. However, as far as the time consumptions of the two algorithms in the same data capacity are concerned, their gap is not very large, and with the increase of the data capacity (number of users), the time rinsing trends are similar that are very much smaller than the linear growth rate. Therefore, the PR4MB algorithm is a higher cost-effective algorithm for the mining Micro-Blogs, since it can get higher analyzing result with a very less time increasing scale than the PageRank algorithm doing.

## 5 CONCLUSION

In this paper, a new weighted PageRank algorithm, called PR4MB, is proposed, which is more suitable for mining Micro-Blogs than the traditional PageRank model. The PR4MB algorithm considers dynamically online behaviours of users as well as links of users, so it can make the mining results more objective and accurate. By computing the activity, quality and credibility of a user, its online behaviours can be measured and the weights of the users are obtained, and so the Google Matrix of the traditional PageRank model is modified into the Weighted Google Matrix. Experimental results show that, compared to the traditional PageRank algorithm, the PR4MB algorithm can get more accurate evaluation results on Twitter with a lighter time cost.

## ACKNOWLEDGMENT

We are deeply indebted to the NSFC (National Science Foundation of China), for this work was supported in part by the NSFC 61273293. Also, this work is partly supported by the CUFE discipline construction.

## References

Page L, Brin S and Motwani R (1999). The PageRank citation ranking: Bringing order to the web. Tech. Report of the Stanford University, USA.

Tunkelang D (2009). A twitter analog to pagerank. The Noisy Channel. Available from http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank.

Weng J, Lim E-P and Jiang J (2010). TwitterRank. In Proceedings of the third ACM International Conference on Web Search and Data mining, New York, USA: ACM Press: 261-270.

Weng J, Yao, Y, Leonardi, E and Lee, F (2011). Event detection in Twitter. Tech. Report of HP Laboratories, USA.

Yin H J (2014). Research on local interested community in large-scale social networks. Dissertation of Ph D, University of Science and Technology of China, P R China.

Kempe D, Kleinberg J and Tardos É (2003). Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD In Proceedings of International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA: 37-146.

Xiong X B (2014). Research on key issues of spreading behavior in microblogging network. Dissertation of Ph D, PLA Information Engineering University, P R China.

Ding Z Y, Jia Y and Zhou B (2014). Survey of data mining for microblogs. Journal of Computer Research and Development, 51(4): 691-704.

Noordhuis P, Heijkoop M and Lazovik A (2010). Mining Twitter in the cloud: A case study. In Proceedings of the IEEE 3rd International Conference on Cloud Computing (CLOUD), Hong Kong, China: 107–114.

Rogers I (2002). The Google Pagerank algorithm and how it works. Available from http://www.sirgroane.net/google-page-rank.

Lin Y, Shi X and Wei Y (2009). On computing PageRank via lumping the Google matrix. Journal of Computational and Applied Mathematics, 224(2): 702–708

Yin D, Hong L and Xiong X (2011). Link formation analysis in microblogs. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information, New York, USA: 1235-1241.

Java A, Kolari P and Finin T (2006). Modeling the spread of influence on the blogosphere. In Proceedings of the 15th International World Wide Web Conference, Edinburgh, England: 22-26.

Cha M, Haddai H and Benevenuto F (2010). Measuring user influence in Twitter : The Million Follower Fallacy. InProceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, USA: 10–17