Improved PageRank Algorithm Combined User Behavior with Topic Similarity

Hao-dong ZHU¹,²,* and Bao-feng HE¹

¹Department of Computer Science, Sias International University, Xinzheng, Henan, 451150, China
²School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, 450002, China

*Corresponding author

Keywords: PageRank algorithm, Similarity weight, PR value, User behavior.

Abstract. Aiming at shortcomings of the drifting theme and the splitting page weight of traditional PageRank algorithm, an improved PageRank algorithm combined user behavior with topic similarity was proposed. Combined with the time factor, the proposed algorithm comprehensively analyzed retweet, comment and mention, measured contribution of three kinds of different behaviors for micro-blog user influence by means of statistical analysis, and used improved TF-IDF calculate the weight of subject relevance to choose pages with higher correlation degree and obtain different PR value. The simulation results show that compared with Micro-blog’s common sorting algorithms, the improved PageRank algorithm has better sorting effect.

Introduction

With the rapid development of Internet technology, especially the gradual maturity of search engines, social networks have become an important platform for the exchange of users online. The size of Chinese Internet users have been expanding, and by the end of 2015, it had reached 6.88 billion. In the face of massive data, the user can quickly and accurately obtain high quality information according to their own needs, which becomes a serious challenge to the search engine technology.

The ideal search engine requirements that it can provide users with the behavior of the user [1] related query information, which requires the existing search engine technology continues to mature. PageRank algorithm is one of commonly used algorithm in the search engine, however, the classical PageRank algorithm [2] ignores the user's individual needs. The content and structure of the network are constantly changing, and users often ignore the time factor, which is not consistent with the law of user behavior. How to get the users’ actual needs and the similarity between the users’ search topics and other factors together? It has become the focus of domestic and foreign that the researchers study the problems.

In PageRank algorithm, all the web pages of the PR value obtained by off-line calculation, which effectively reduces the computation of online query and greatly reduces the query response time. However, on the one hand, people's query has the theme feature and PageRank ignored the relevance of the topic, leading to the results of the correlation and the theme of the reduction; on the other hand, the old page level will be higher than the new page. Even a very good new page will not have a lot of upstream links, there are a series of problems, such as the theme of the drift, ignoring the user link quality issues, the emphasis on old web [3] issues and so on. In view of this status, this paper presented an improved PageRank algorithm which based on web page quality, time factor updating, user personalized search [4] and web pages similar theme [5].
Relevant Basic Knowledge

PageRank Algorithm

It will divide a certain weight according to the importance of the page and transfer to the search page, so as to determine the search page using PageRank value (referred to as the PR value)[6]. Finally, it judges the importance according to the web page of the PR value level, improving the quality of the user to quickly search[7].

The researchers generally believe that the link between the web page is established by hyperlink, then the number of pages constitute a super large in the end. Users randomly select a web page to browse from all web pages. There are two options for each web page by using hyperlinks on the web page[8]: ①To this end. ②Continue to select a link to browse. Therefore, we calculate of a web page, for example the rank of web page A that the standard formula as the standard (1).

\[ PR(A) = (1-d) + d \sum_{i=1}^{n} \frac{PR(P_i)}{C(P_i)} \]  

(1)

In formula (1), \( n \) represents the total number of pages, \( PR(A) \) represents a web page (e.g., page A) \( PR \) of all the pages; \( PR(P_i) \) represents the value of the page A page of the link to the \( P_i \), \( C(P_i) \) represents the number of outbound links page \( P_i \); In order to avoid the emergence of the link between the web and the steady state probability cannot converge, here, the introduction of damping coefficient \( d \) (\( 0<d<1 \)), the value of 0.85, it means that users continue to visit the current page probability. \( 1-d \) represents the probability that a user will randomly access other web pages from the current web page. According to the traditional algorithm, \( PR \) value a page of \( i \) is only related to its penetration.

Random Walk Model

It converts the behavior of users browse the web page into a web link structure, in the structure of the link according to the direction of random forward, and with a certain probability of random jump to other pages[9]. The probability that each page will be linked to the user will tend to a stable value after a number of visits. Because the user randomly selects one of the links at the same time as the equal probability in the current page to browse and repeats the process continuously. The iterative relation of the algorithm is shown in the formula (2).

\[ PR(i) = \frac{1-d}{n} + d \sum_{j \in \text{in}(i)} \frac{PR(j)}{\text{out}(j)} \]  

(2)

In formula (2), \( i \) represents a collection of pages that point to a web page \( I \), \( \text{out}(j) \) represents a collection of pages that point to a web page \( j \).

Haveliwal of the Stanford University in 2002 in Topic-sensitive pagerank proposed PersonalRank algorithm[10]. We replace (2) in the page nodes with the user and item that we can calculate in the global for each user and item of importance, and thus give the overall ranking. In fact, we need to calculate the user behavior node which is relative to a user node m correlation. This algorithm can be used to sort for all relevant user personalization behaviors. The iterative relation of the algorithm is shown in the formula (3).

\[ PR(i) = (1-d)ri + d \sum_{j \in \text{in}(i)} \frac{PR(j)}{\text{out}(i)} \], \( r_i = \begin{cases} 1 & i = m \\ 0 & i \neq m \end{cases} \]  

(3)

In formula (3), \( r_i \) is replaced by \( 1/n \), which means that the probability of starting from different points is different. \( m \) represents the target user node we recommend. Using (3) calculation, the correlation degree of all vertexes with respect to the target point \( m \) is obtained.

For PageRank, since the initial access probability of each node is the same, the initial access probability of all nodes is \( 1/n \) (\( n \) is the total number of nodes). Although the random walk model takes the actual needs of users into account, it ignores the relevance of the content of the page.
Analysis of Current Correlation Improved Algoritms

G Kleinberg, J Teunte et al.[6] focused on the analysis of the propagation properties of Twitter, through the comparative analysis of a large number of Twitter data, the study of the impact of the user in the spread of certain topics. The results show that, in the micro-blog retweet and quotation, there is no direct relationship on penetration of distribution and user influence order. Conversely, influential users play an important role in a variety of topics. Chen Hongchang et al.[7] focused on the analysis of retweet, comment and three kinds of acts mentioned in micro-blog, using statistical methods to enumerate users who are in these three aspects related to the number of carries on the contrast analysis, and measured the level of the influence of communication users, finally, they put forward a kind of based on behavior of weights allocation algorithm; Lu Gang, Li Gengsheng et al.[8] proposed an improved algorithm which was in the traditional algorithm based on into the link similarity between web pages and referenced page release time, improving the precision and recall of search results, however, now it seems that they ignore the similarity between users of the factors.

Because of the lack of the influence of the traditional page ranking algorithm on the sorting results, T. Havdiwala [9] et al. improved in this area, which not only has the network topology as a metric, but also has the influence factors such as context correlation and subject sensitivity. Finally, through a large number of simulation experiments proved their assumptions. Weng et al. [10] proposed the TRank (TwitterRank) algorithm, which not only considers the factors such as network structure and user interaction, but also analyzed the similarity of the subject. Eventually, a combination of all these factors was calculated and the influence of each twitter user. Xu Zhiming, Li Dong et al. [11] defined the strength of the user relationship as the similarity between the users, and proposes a method for computing the similarity of the users’ attribute information. However, these researchers don’t combine the user browsing behavior characteristic with the similarity of theme of the page to research.

Improved PageRank Algorithm of This Paper (BSPR)

Based on the time factor analysis, this paper uses the improved space vector model to calculate the relative weight of the subject. The improved algorithm extract the relevant high degree of web pages allow people to freely choose, finally making us obtain the corresponding PR. Therefore, the problems that Theme drift, emphasis on old web pages, web weight sharing are suppressed to a certain extent. Thus, this paper proposes BSPR (based on Behavior and Similar Theme PageRank) algorithm, the formula such as the formula (4).

\[
BSPR(u_i) = (1 - d) + d \times (\alpha \sum_{v_j \in U_{ui}} B_w(u_i, v_j) \times BSPR(v_j) + \beta Wu_i + \gamma W_t)
\]

(4)

Among them: \(u_i\) and \(v_i\) for micro-blog users, \(F_{\beta (ui)}\) represents the final set of fans \(u_i\) fans, BSPR (\(u_i\)) on behalf of the user \(u_i\) PR, \(BSPR(v_j)\) on behalf of the user \(v_j\) PR, \(BW (u_i, v_j)\) as the scale factor of \(u_i\) page weight, \(v_j\) user, if \(v_j\) users pay attention to page \(u_i\) for retweeting, interactive topic and micro-blogging mentioned behavior, and scaling factor working, \(W_t\) represents the time feedback factor of the web page vi. \(\alpha, \beta, \gamma\) represent user behavior, topic similarity and time factor accounted for in the calculation of the proportion of the PR value., And to make \(\alpha + \beta + \gamma = 1\), so as to more accurately estimate the value of BSPR, making micro-blog’s influence on the order of the user is more true. The initial PR value of each user is set to 1.

User Behavior Analysis

In micro-blog, there are many factors affecting the ranking of the user influence, but the users’ behavior exists based on micro-blog retweet, comments and mention[11].

Three different user behavior on the role of micro-blog user influence ranking is different, which is based on this constructing scale factors. In this paper, we use the method of constructing a
directed weight network, which is forwarded by the user retweet, the user comments, micro-blog mentioned, and the distribution factor is given in turn, then the weight formula is constructed as (5).

\[ B_w(u_i, v_j) = \eta R_{ji} + \theta C_{ji} + \omega M_{ji} \]

(5)

In (5) the formula, \( R_{ji} \) represents micro-blog forward weight, \( C_{ji} \) represents micro-blog comments, \( M_{ji} \) represents micro-blog referred to the weight. In the micro-blog network, when multiple networks need to be combined, the direct sum of the simple edge weights is not reasonable. Thus the introduction of \( \eta, \theta, \omega \) distribution factor, which respectively, represents the user retweet, user comments, micro-blog mentioned three kinds of behavior contribution in the page ranking algorithm.

In general, we adopt the measure standard that the users publish the micro-blog to statistic the impact of the crowd, so as to get the formula (6).

\[ \eta = \frac{R_{ni}}{R_{ni} + C_{ni} + M_{ni}} \]

(6)

In the formula (6), \( R_{ni} \) represents the forwarding impact, \( C_{ni} \) represents the commented on the impact of people, \( M_{ni} \) represents the mention the impact of people. Thus, the values of \( \theta \) and \( \omega \) can be obtained in turn. The formula (5) is obtained by normalizing the formula (7).

\[ B_w(u_i, v_j) = \frac{W(u_i, v_j)}{\sum_{k=1}^{n} W(u_k, v_j)} \]

(7)

In the formula (7), \( B_w(u_i, v_j) \) on behalf of micro-blog users paying attention to the contribution size of the user \( u_i \), \( n \) represents the total number of users \( v_j \) concerned about the object. \( k \) represents micro-blog retweet the case several times. After the treatment \( \alpha, \beta, \gamma \) will fall between \([0,1]\). Then we make full use of scale factor and the formula (4) to calculate all users of the PR value.

**Topic Similarity Analysis**

In the improved algorithm based on PageRank, we consider only the users’ behavioral factors, neglecting correlation between users, which in some extent solves the topic drift phenomenon. However, problems of PageRank algorithm itself did not really solved. Li Zhiyeng, Yang Wu and others [12] use the improved space vector model, that is, TF-IDF formula, respectively calculating weights \( m \) and \( n \) on the terms of the \( t_i \) where is a corresponding link between the user \( u \) and user \( v \). Defining Micro-blog content as the vector space model in the vector, the paper uses the cosine vector measurement method, so as to calculate the similarity of the contents of the micro-blog. Therefore, the formula for calculating the similarity of the similarity is shown in (8).

\[ W(u_i, v_j) = \frac{\sum_{i=1}^{n} W'(i, u) \times W'(i, v)}{\sqrt{\sum_{j=1}^{n} W'^2(j, u) \times W'^2(j, v)}} \]

(8)

In formula (8), \( W(u_i, v_j) \) expresses micro-blog user \( u_i \) and micro-blog user \( v_j \) content similarity value, \( W'(i, u) \) expresses the word \( t_i \) in the micro-blog user \( u \) in terms of the right value, \( W'(i, v) \) represents the term \( t_i \) in micro-blog users.

We use the formula (9), to calculate proportion of the similarity of the page content in the formula (4).

\[ W_{u,v} = W(u_i, v_j)/\sum_{k \in O_u} W(u_i, k) \]

(9)
In the formula (9), \( W_{u,v} \) represents the weight of the content similarity of all the degrees of the user \( v \) in the user \( u_i, O_{u_i} \) represents a collection of user \( u_i \) pointing to the user \( v_i, k \in O_{u_i}. \)

**Time Feedback Factor**

In general, the number of web pages being searched by the search engine to indicate the length of time that the web page itself exists, In view of the existing problems which lay particular stress on old web pages and ignore new valuable web pages in the PageRank algorithm, the paper puts forward the time feedback factor to revise the formula (4), the formula as (10).

\[
W_t = e^T
\]  

(10)

In the formula (10), \( T \) represents the number of times that the web page is searched by the search engine. \( w_t \) as a web page time feedback factor. \( E \) as a constant, it is the value of \( e=(1-d)/n \), which is related to the total number of \( n \) that search engine access to the project, but also affected by the standard formula (1) the impact of \( d \). Generally speaking, the search engine's search cycle is 15 days to 30 days. If a web page exists in the network for a long time, accordingly, it is in each search cycle to be accessed to the greater the probability. In other words, single page exist time is proportional to the number of search engine access.

From the improved PageRank algorithm to modify the formula we know that if the larger number of twitter users, corresponding to, the user retweet, comment, mention to the higher the number. The user behavior related factors accounted for the greater proportion when we calculate the PR value, therefore, the introduction of correlative coefficient \( \eta, \theta, \omega \) is particularly important. The greater the total number of micro-blog, the more interaction between the user, the higher the similarity between the pages, to a certain extent, it avoids the problem of topic drift. The algorithm introduces the time feedback factor, which overcomes the difficulty of extracting the time of the web page and can judge the date of the web page more accurately. The size of the PR value is affected by the length of time the web publishing, which is conducive the old web page precipitating and the new page rapidly rising.

**Improved PageRank Algorithm Description**

According to the calculation method of the BSPR algorithm proposed in this paper, the main steps are as follows:

1. Calculate each micro-blog users page in-degree and out-degree number.
2. Get the last modification time of each page.
3. Construct the network vector by using the points, edges and weights of the network.
4. Statistics the crowd affected by micro-blog and standardizes processing user behavior weight.
5. Calculate the proportion of the weight of the page similarity by using the cosine vector measurement method.
6. Calculate the time feedback factor of the web page, according to the cycle times \( T \).
7. By repeated experiments, the balance coefficient \( \alpha, \beta, \gamma \) in the proportion of the whole formula, according to the formula (4) to calculate the \( BSPR(u_i) \), so as to get the ranking of micro-blog's user influence.

This paper focuses on the calculation of the influence of micro-blog users (Step4, Step5, Step6 are the core of the paper steps).

**Experiment Simulation**

**Experimental Preparation**

This paper takes Sina micro-blog from June 3 to June 20 in 2015 as the experimental data. Part of the user micro-blog information is taken as research data, including the user set \( U \), micro-blog data
sets \( Z \), micro-blog web pages \( \theta_j \), micro-blog text sets \( X \) and other. The main contents of the data storage:

1) User information: user uid, user name, gender, location, create micro-blog time, the user home URL, fan number, the number of concerns, micro-blog, the collection number.

2) Micro-blog information: micro-blog mid, publishing time, micro-blog content, micro-blog sources, forwarding number, comments, was praised, published by the user uid, micro-blog belongs to the subject.

3) Listening relationship: user uid, fan id.

4) Concern: the user uid, concerned about the id.

Experiment environment settings: 1) hardware platform: 1 HP servers, Intel i7-4790, 1TB 3.4GHz hard disk and 2 HP 3.6GHz Intel PC processor, 4G memory, 500G hard disk. 2) software platform: Win7 operating system, MATLAB simulation tool, MySQL 5.5 database.

Due to the restrictions of micro-blog Sina, each user can only get to 200 people at most, so the collection of good friends is not very comprehensive. We merge all of the collected micro-blog user data, and simple clean it with the crawler tool to remove duplicate entries. The final user data statistics, such as table 1.

Table 1. User data statistics.

<table>
<thead>
<tr>
<th>Project data</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>The total number of users</td>
<td>10413</td>
</tr>
<tr>
<td>The total number of micro-blog</td>
<td>84,168</td>
</tr>
<tr>
<td>Micro-blog retweet</td>
<td>27,759</td>
</tr>
<tr>
<td>The user's friends</td>
<td>1,391,718</td>
</tr>
</tbody>
</table>

Labeling, extraction of the collected data, we get the number of edges and average degree and average in degree related data, which is based on retweets, user reviews, users refer to the number of nodes, as shown in table 2. According to the formula (9), Calculate the page similarity proportion of micro-blog users in the formula (4), and the results will be calculated as a factor of the users’ influence ranking.

Table 2. Micro-blog basic data.

<table>
<thead>
<tr>
<th>micro-blog</th>
<th>The number of nodes</th>
<th>The number of edges</th>
<th>The average out degree</th>
<th>The average In degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>retweet</td>
<td>10413</td>
<td>90132</td>
<td>18.31</td>
<td>19.13</td>
</tr>
<tr>
<td>comment</td>
<td>10413</td>
<td>117846</td>
<td>31.23</td>
<td>32.46</td>
</tr>
<tr>
<td>mention</td>
<td>10413</td>
<td>31843</td>
<td>7.33</td>
<td>5.22</td>
</tr>
</tbody>
</table>

Experimental Results and Analysis

Firstly, the probability value of each node is initialized, the initial access probability of the node corresponding to the \( m \) is 1, and the initial access probability of the other nodes is 0. After data calculation, this paper in the formula (5) [8], takes \( \eta = 4/7, \theta = 2/7, \omega = 1/7 \).

Secondly, after repeated experimental simulation, we calculate \( \alpha, \beta, \gamma \) three scaling factor values in the BSPR algorithm, the results shown in figure 1.

From figure 1, we know when \( \alpha, \beta, \gamma \) take different values, BSPR algorithm of the PR value with the change, in order to illustrate the influence of \( \alpha, \beta, \gamma \) on the improved PageRank. When \( \alpha = 0.3, \beta = 0.6 \) or \( \alpha = 0.6, \beta = 0.3 \), user behavior and topic similarity factors accounted for a large proportion and the time factor smaller proportion, leading to the micro-blog page of the PR value is generally low, the PR value difference is smaller; When \( \alpha = 0.2, \beta = 0.4, \gamma = 0.4 \), user behavior factors are the proportion of smaller than page topic similarity and time factor, the page the PR value gap increases. With the time factor of \( \gamma \) increasing, the recently updated web page of the PR value rises. New web pages floating, old pages sinking. Therefore, the proportion of the algorithm
is set to $\alpha = 0.2$, $\beta = 0.4$, $\gamma = 0.4$, which reduces the user theme drift phenomenon, makes up for the lack of weight sharing page, more conducive to the user's influence ranking.

Finally, in order to verify the effect of the article BSPR algorithm on the micro-blog user influence ranking, this paper compares it with the three methods commonly used in micro-blog. The simulation results are shown in figure 2.

![Figure 1. Values of $\alpha, \beta, \gamma$ analysis chart.](image1.png)

![Figure 2. Comparison results of five algorithms.](image2.png)

In figure 2, the horizontal coordinates, in the same query conditions, the query topK% micro-blog users. The vertical coordinates, user influence coverage ratio. From figure 3, we know that the BSPR algorithm is superior than BWPR algorithm(User behavior through algorithm, retweets, comments, and other factors mentioned in the analysis of user, to calculate the influence order of users), WPR algorithm (The theme correlation algorithm, by constructing a link analysis of micro-blog users / web content based on user influence ranking) and TURank algorithm (Combined with time factor algorithm, the time factor is introduced to observe the influence of users in a certain period of time). It shows that the micro-blog data information filtering and user related behavior, the theme of similarity, the combination of the time factor and as the basis for the weight distribution is more close to the reality, and achieve a good effect.

Conclusions
Aiming at the shortcomings of the traditional PageRank algorithm, the drifting theme, splitting page weight, new web links less, the user behavior, the theme of the relevant weight, time factor and other factors are introduced into the PageRank algorithm, then an improved PageRank algorithm based on the users’ behavior and topic similarity is proposed, and the algorithm is applied to the ranking of micro-blog users' influence. Using the collected data to simulate the experiment, the improved algorithm can get a better sorting effect, which shows the feasibility and superiority of the BSPR algorithm.

Acknowledgement
This work was supported in part by the Science and Technology Plan Projects of Henan Province of China under grant No. 152102210357 and No. 152102210149.

References


