

**Study on Improved Influence Page Rank Algorithm Based On User Behavior**

Yun Wu

**ABSTRACT**

User influence ranking is an important subject of social network research, for public opinion control, important dissemination and so on is a great significance. The traditional PageRank algorithm only considers the relationship between users, this paper combines the Page Rank algorithm with the user behavior, integrates the follower quality, comment rate, forwarding rate and whether the authentication behavior and other user behavior factors, proposed an improved algorithm URank. Experiments based on the SIR propagation model show that the URank algorithm is superior to the PageRank algorithm in calculating the accuracy of the influence index for large-scale data sets.

**INTRODUCTION**

In the social network, the influence\[1\] high users play a vital role in the dissemination of information, through the rankings to find one of the most influential nodes or nodes set, which can effectively control the information transmission, optimize the available Resource scheduling, conducive to public opinion control and other research work carried out.

According to the research situation of scholars at home and abroad can be user influence calculation method can be summarized as the following: Based on PageRank\[2\] algorithm to calculate user influence, Calculate user influence based

Yun Wu, BaoTou Teacher's College, China
on user behavior, Calculate User influence Based on Information dissemination, URL tracking based on the way to calculate the user influence. Meeyoung Cha [3-4] and others based on user behavior on the user behavior to influence the calculation, and then calculate the results through the relevant mathematical formula Calculation, and ultimately get the user’s influence index. Based on URL tracking, Eytan Bakshy[5] and others began to calculate from the key influencing user nodes to study the forwarding situation of all followers to study the propagation situation until the end of the transmission, and finally get the influence index of the user node. Unkelang[6] proposed an improved PageRank algorithm for the social network node influence sorting algorithm—TunkRank algorithm by calculating the influence of the followers to sum up the impact of each user to follow the index. The greater the influence index, the greater the probability that the user’s message will be read by the follower. However, the algorithm is also inadequate, the experiment shows that the number of followers is not necessarily transformed into influence, some have a high followers of the user, through the algorithm to calculate the influence index is not high.

PageRank algorithm based on user influence assessment method, that is, from the user to assess the influence of followers, such method is simple, practical, users have been influential rankings of the general trend on the whole. However, the concern is only the relationship between the nodes, taking into account the less influential factors. In this paper, we propose an improved algorithm, Urank, which combines the PageRank algorithm with the user behavior to calculate the user influence index, which is based on the user behavior to calculate the user’s influence.

**URANK ALGORITHM**

**Follower Quality**

Using the PageRank algorithm to calculate the user influence index[7-8] that followers quality. The follower’s contribution value is expressed, that is, the user’s follow we quality index, which is calculated as follows:

\[
URank_F(m) = (1 - d) + d \sum_{n \in Z(m)} \frac{URank(n)}{Link(n)}
\]

Where Z (m) represents the set of followers of user m, user n is a follower of user m, represents the quality index of user n, represents the total number of user n followers, d is the damping coefficient, Reference[9] and past experimental experience, in this article default to take d = 0.85, which means that the user random follow the probability of other users.
Implementation of URank Algorithm

The improve algorithm is based on the quality of the user and the quality of the follower, and the followers are used to achieve the contribution of the followers. Therefore, the linear integration method is used to combine the user’s own quality index with the follower’s quality index. The method of calculating the user influence index, URank algorithm core calculation formula is as follows:

\[ URank(m) = \lambda \times URank_S(m) + \mu \times URank_F(m) \]  \hspace{1cm} (2)

Which \( URank(m) \) represents the URank value of the user m, that is, the influence index, \( URank_S(m) \), which represents the user’s own quality index, \( URank_F(m) \), which represents the URank value of the user m, that is, the quality index of the user m follower. \( \lambda, \mu \), Respectively, said the quality of the user’s own quality and followers of the weight. The improved algorithm not only takes into account the quality of the user itself and takes into account the impact of the follower on the user, adding the weight to make the result of the increase is reasonable. The question of weighting is summarized in the third section of this paper by experimental calculation. The formula(6) is further decomposed to obtain:

\[ URank_F(m) = (1-d) + d \sum_{n \in Z(m)} \frac{URank_F(n)}{Link(n)} + d \sum_{n \in Z(m)} \frac{URank_S(n)}{Link(n)} \]  \hspace{1cm} (3)

From the formuals (1) and (2), \( d \sum_{n \in Z(m)} \frac{URank_S(n)}{Link(n)} \) is value. It is known that PageRank converges to a constant, then the formula (3) also converges to a constant, so that the URK algorithm core formula (3) converges to a constant. So the improve algorithm through the iterative calculation of the user influence index converges to a constant, the resulting constant is the user’s comprehensive influence index.

EXPERIMENT AND ANALYSIS

This experiment was conducted using the data set in the real network. The experimental data set was derived from the Twitter data set in the SNAP project at Stanford University. The resulting Twitter data set nodes were infected with the SIR [10] model. The SIR propagation model takes one of the nodes in the initial phase as the infection state, and all other nodes in the data set are also sensitive. When the infected node and the recovery node reaches the set value, it is the
infection ability that is initially set to the infected state node, that is, the influence. The infection ability of each node in the data set was obtained, and the infection experiments of all nodes were carried out 100 times. Calculate the average of the infectivity of each node, and then use the average set of all nodes as the baseline data for subsequent experiments.

**Get the Optimal Weight**

Based on the previous experimental experience and the social network reality, the weights in the URank algorithm are selected, and the different weights after the assignment are composed of a weight group, and the optimal weight is judged according to the hit rate. [11-12] The hit rate calculation process is: First, the data set is calculated by the influence algorithm, and then the proportion of the n nodes in the first n nodes of the reference data is recorded. This ratio is the hit rate of the algorithm. Figure 1 depicts the top five curves with relatively good hit rates for different weight groups. $\alpha$, $\beta$, $\gamma$ that the weight of the three kinds of user behavior.

![Figure 1. Comparison of the results of different weight groups.](image)

As can be seen from Figure 1, the range of hit rates of the URank algorithm obtained by the different weight groups is different, the overall hit rate is
maintained between 0.79 and 0.92, when the value group is $\lambda = 0.6$, $\mu = 0.4$, $\alpha = 0.5$, $\beta = 0.2$, $\gamma = 0.3$, the algorithm hit rate range between 0.87 and 0.92, the hit rate range is better than other weight groups get the hit rate range, while the hit rate trend is better than other weight groups of the hit rate trend, so this set of parameter s as the optimal weight group. From the results of the optimal weight group, we can see that the quality of the user is the key of its influence, and it is also affected by the quality of the follower. From the value of the three user behavior values can also be seen that the majority of user message forwarding number is greater than the number of comments, whether the certification impact on the user there is a greater impact on social networks that have a greater influence on the user must Most of them are real-name authentication users, from the experimental results can be seen in line with this feature.

**Comparison of Accuracy**

With the hit rate as the indicator of the accuracy of the decision, the higher the hit rate that the algorithm more accurate sort. In order to verify the accuracy of the improved URank algorithm in calculating the influence index of the user node, the results of the PageRank algorithm and the hit rate experiment based on the URL tracking algorithm are used as the control. URank algorithm weight value from the last optimal weight of the experimental results, the application of the data set and the baseline data to get the following results.

![Figure 2. Data set 1 hit rate comparison.](image)

As shown in Figure 2, in the data set 1, as the number of nodes increases, the hit rate of each algorithm is slightly improved. In the hit rate test with nodes less
than 800, the URank algorithm has a hit rate less than the PageRank algorithm and slightly higher than the URL tracing algorithm. In the experiment with nodes higher than 800, the URank algorithm has better hit rate than the other two algorithms. The third section of this article knows that the influence index of the URank algorithm is determined by the user’s own quality and followers’ contributions. The PageRank algorithm only considers the user’s own quality, combined with the experimental results and the URank algorithm formula to introduce the contribution of the followers to the user when the number of nodes is less than 800 is not fully reflected, and in the number of nodes greater than 800 when the contribution of the followers increased significantly, URank algorithm hit rate is better than the other two algorithms. It can be concluded that the URank algorithm is more suitable for dealing with large data sets with nodes greater than 800, and the PageRank algorithm is more suitable for processing small data less than 800.

![Hit Rate vs. Number of Nodes](image)

Figure 3. Comparison of the results of the data set 2 hit rate.

In the hit rate test of data set 2 (Figure 3), the large data nodes with nodes ranging from 1000 to 10000 are selected. With the increase of the number of nodes, the hit rate of each algorithm is slightly increased and decreased. This is because the number of zombie nodes in the number of adjacent nodes in the experiment is different, and the randomness of the experiment is different. Down trend, but the overall three algorithm hit rate stabilized. From the experimental results in Figure 3 can be clearly seen, the use of URank algorithm to get the overall rate of hit better than the use of PageRank algorithm and URL tracking.
algorithm to get the hit rate, when using 10000 nodes for hit rate experiment, Urank algorithm can find The node node is 92% of the nodes, and the PageRank algorithm and the URL tracing algorithm can only find 87% and 88% of the nodes in the node set. We can conclude that the URank algorithm has a higher accuracy in the user influence index than the PageRank algorithm and the URL tracing algorithm in dealing with nodes with a number greater than 1000. The URank algorithm has obvious accuracy in calculating the influence index of social network nodes in large data sets. The above results show that it is feasible to combine the PageRank algorithm with the user behavior to calculate the user influence index, and the URank algorithm is superior to the PageRank algorithm and URL tracing in calculating the reliability index of the large data set nodes. algorithm. So the URank algorithm is comprehensive and more realistic, better reflect the user’s comprehensive influence.

CONCLUSIONS

The URK algorithm is based on the PageRank algorithm and the user behavior. The improved algorithm is composed of two parts: the user’s forwarding rate, the user’s comment rate and the user authentication. The quality of the followers of the four assessment conditions, and assigned to different user behavior with different weights, and finally the proposed algorithm for the results of the improved analysis to get the optimal weight group, and through the size of two data Set the algorithm to verify the URank algorithm in the calculation of social network users influence index more accurate and comprehensive, more in line with the reality of the conclusions. The next step will continue to look for other factors related to the user’s influence in order to obtain more comprehensive and accurate results.

REFERENCES


9. Chen Hao, Based on the microblog user influence ranking algorithm based on Hadoop [D]. East China University of Science and Technology 2014.(In Chinese)

