Identify important users on social networks–Cases from Sina Weibo

S Ma¹, Y H Ma¹, *
¹Business School, Shandong Normal University, Jinan, 250014, China
Corresponding author: Y H Ma (e-mail: Yhma@sdnu.edu.cn)

Abstract. In recent years, more and more studies on the strength of interpersonal relationship show that there are some positive social connections in online social networks. In addition, in the process of information dissemination, there may be some key nodes. In order to determine the key nodes of online social networks, we propose a method of edge weight recognition based on historical interaction records. According to the forwarding quantity, the influence of nodes is quantified, and the top node is selected as the key node. We apply our method to a real network and get good results.

1. Introduction
The emergence of online social platforms (such as microblog, twitter, Douban, Facebook) has changed the way we get information, and caused a major reform of the media industry. The information on online social platform is diverse, the amount is huge, and the platform has an extremely complex structure, which makes the diffusion mechanism of information on social platform more complex. However, the final information usually spreads along some specific channels. For example, many users on online social platforms like to forward information from their friends' microblog space [1], which makes the information spread. However, there are some users on the online social platform who do not spread information from their friends based on Forwarding behavior. They choose to write their own microblog after observing the microblog published by their friends, so as to spread the information. The following relationship between users on the online social platform affects the diffusion of information. There are some "leaders" on the platform who can lead the information. They often have a large number of "fans". The participation of such "leaders" can speed up the diffusion of information on the network platform. Therefore, identifying these key nodes from social networks will help to control the diffusion of information, and it is possible to further control the diffusion of information.

Information is transmitted through the Internet. The premise of studying information dissemination is to analyze the network characteristics and network structure. The research work of network information dissemination usually draws lessons from the infectious disease model, and uses SIS [2] model and SIR [3] model to simulate the information dissemination. Zanette et al. [4] established the mean field equation to study the propagation mechanism of rumors in small world networks. The experimental results show that the maximum range of rumors can be spread will be limited. Guille et al. Use machine learning method to predict the time trend of information diffusion by analyzing the three dimensions of semantic, social interaction and time based on Twitter users' attention to the relational network [5]. LV pointed out that information dissemination is different from disease dissemination and has the characteristics of memory, that is, the past interaction between users can affect the current information dissemination [6]. When Yang et al. Analyzed the dynamics of information transmission
on twitter, they found that the interactive network composed of communication behaviors among users is more valuable than the network concerned by users [7]. Boyd et al. Based on the historical forwarding data of Twitter users, analyzed the motivation of users' forwarding behavior and the content topic tendency of the forwarded information [8,9]. Kahanda et al. Believe that as a supervised learning method, manual annotation of training data costs a lot, and prove that the historical interaction information of users in social networks can effectively identify the strong and weak relationship between users [10].

To sum up, many literatures have emphasized the influence of user's historical interaction behavior on current user's communication behavior, and used historical interaction data for analysis and prediction. At the same time, some scholars have expounded the effectiveness of using historical interaction information to describe relationship strength.

2. Materials and Methods

2.1. Users follow the network

In most online social networks, users usually get more information by "following" other users. Taking microblog network platform as an example, users are divided into two roles: "follower" and "followee". The following relationship between users indicates the direction of information dissemination. The network formed by this "following" relationship between users is a directed complex network. As shown in figure 1, if user B follows user a, it can be considered that user B is the follower of user a. When user a publishes information, the information is propagated from "follower" a to "followee" B; when user B publishes information, the information will not be propagated to user a, which also means that the direction of information propagation is opposite to that of following relationship. If user B pays attention to user a, then the information is propagated from user a to user B. Other paragraphs are indented (Body text Indented style).

![Figure 1 An example of user following relationship in social network.](image)

This model will measure the strength of the relationship in the network based on the historical interaction data between users. Taking microblog network as an example, the historical interaction data includes the comment frequency and forwarding frequency between users. This paper selects the historical forwarding frequency between network users as an index to measure the strength of the relationship between users. Although the reading frequency also reflects the intensity of netizens' opinion interaction to a certain extent, it does not play a direct role in the actual diffusion of a microblog in the network. On the contrary, the user's microblog forwarding behavior directly leads to the rapid spread of microblog information in the network. A user's forwarding behavior will enable all his followers to browse the microblog information.

In this paper, a directed network is constructed according to the following relationship between users, and then a user following network G is constructed based on the forwarding records of all users in a certain period of time in the past. The nodes in the user following network are network users, and the connecting edge between nodes represents the following relationship between network users. The forwarding records between users in a certain period in the past are defined as the weight of the edge, which is recorded as:

\[ G = (V, E, W) \]
Among them, V represents the user set, e represents the following relationship between users, and W represents the weight value of connecting edges between users. Figure 2 shows an example of the user following the network. The right figure is the network constructed according to the left figure, and the five pointed star represents the network user. If there is a following relationship between two users, and they have forwarding records in a certain period of time in the past, the two pentagons are connected by an edge, and finally a user following network is constructed.

Figure 2 An example of the user following network

2.2. Identify important users
Some text. In the network, ordinary users usually interact through opinion leaders, who are the influence center of the group. Because in every group, there are some people who are especially good at communication. They are the connection center of the community and a very important part of the social network. They make information penetrate into the network faster. Next, this paper will identify these key nodes.

This section intends to improve PageRank algorithm based on network topology to evaluate the importance of nodes. In this paper, the link from node a to node B is regarded as the vote from a to B, and then the importance of the node is quantified according to the voting source (or even the source of the source). In order to solve the problem that the sorting result is not unique due to the unconnected network, we add a basic node "ground" to the network and connect it with all other nodes in the network in two directions, so that the whole network becomes a strongly connected network and the subsequent calculation is simplified.

In order to quantify the importance of nodes in the network, each node is given the same importance score s (I), and the score value of each node will be assigned to the following nodes, that is, the score value of each node will flow from the follower to the follower along the directed connected edge. That is to say, each node in the network will "absorb" the score of its following nodes, and the more the number of corresponding connecting edges is, the more obvious the "absorption effect" is. As shown in figure 3, user a pays attention to four users (figure i), and its score value s (a) will be allocated to the four users he follows in proportion according to the connecting edge (Figure j).

![Figure 3 Example diagram of node score allocation.](image)

The formula for calculating the score $S(A)$ of node a is as follows:

$$S(A) = \alpha \sum_{Y_i \in X(A)} \frac{S(Y_i)}{n_i} + \frac{(1-\alpha)}{N}$$
Where \( X(A) \) represents the set of all nodes pointing to node \( a \). \( S(Y_i) \) is the score of node \( Y_i \), node \( Y_i \) is a node in the set of all nodes pointing to node \( A \). \( n_i \) is the degree of node \( Y_i \), that is, the number of edges \( Y_i \) points to other nodes. \( N \) is the total number of nodes. \( \alpha \) is the damping coefficient, generally 0.85.

At the beginning of the experiment, the initial score of the nodes in the network is \( \frac{1}{N} \) (where \( N \) is the number of network nodes), and the score of all nodes will be repeatedly calculated in each subsequent calculation, until the score of each node in the network reaches a stable state in a certain calculation. Finally, we rank the score values of all nodes from large to small, and select the node with the highest score value as the key node.

3. Data description
This work is based on Sina Weibo platform, grabs the relevant data about the event of "Changsheng biological vaccine fraud" on the platform, further constructs the user following network, and makes some basic data statistical analysis. Table 1 shows some basic topological parameters of user following network, such as number of network nodes, number of connected edges, average degree, clustering coefficient and average shortest path.

| N      | M   | \(<k>\) | C    | L    |
|--------|-----|---------|------|------|
| 1.143  | 0.285 | 0.286   | 0.286 | 0.593 |

4. Results & Discussion
We apply our method to build a user following network to find important nodes on the network. According to the algorithm, the scores of all nodes in the network are calculated recursively, and it is known that the scores of all nodes converge to a stable state. Figure 4 describes the score distribution of all nodes in the steady state, as shown in the figure, the ordinate represents the node score value, the abscissa represents the node number, the score values of all nodes in the final stable state (from large to small) roughly follow the long tail distribution, and the nodes above the red dotted line in the figure are the key nodes selected in this paper. After statistics, we can find that in the final stable state, the scores of most nodes in the network are about 0.001, and only about 7% of the nodes can reach above 0.003.

![Figure 4 Distribution chart of node score value (descending order).](image)

According to the distribution of node scoring values in figure 4, we select a small number of nodes with the top ranking of scoring values (descending order) as the key nodes in the network, and then we
need to consider a problem: how many key nodes need to be selected at least to make the influence of the selected key nodes cover the whole network as much as possible. In order to solve this problem, we use the total number of fans of the selected node to measure the influence range of the selected key node. On the actual network platform, the "fans" of "leaders" are the direct recipients of their published information. As shown in figure 5, nodes 2, 3, 4 and 5 are followers of node 1, so the influence range of node 1 covers nodes 2, 3, 4 and 5.

Figure 5 Follower of node 1.

Figure 6 shows the influence coverage rate of key nodes when selecting different number of key nodes (ranking nodes in descending order, taking the top part of the score value). The red solid line in the figure is the slow down point of the rising speed of the influence coverage rate curve. As can be seen from the figure below, when the number of selected key nodes is about 18, the rising speed of the influence coverage of the selected key nodes begins to slow down, which means that even if we select more key nodes, the overall influence coverage of the selected nodes will not increase significantly. Our goal is to select as few key nodes as possible to make the total influence coverage of the selected key nodes larger. According to the data in figure 6, the number of selected key nodes should be controlled between 15 and 25. In this paper, we select the top 15 nodes as the key nodes in the network, which accounts for 2.7% of the total number of nodes in the network.

Figure 6 The influence coverage of key nodes with different number of key nodes.

5. Conclusions
How to analyze and identify the key nodes in the network from a global perspective is meaningful. This paper focuses on how to identify key nodes in online social networks, and proposes an effective quantitative method. All the key nodes play an important role in the network, so that different online
communities can exchange information with each other. With the help of the identification of key nodes, we will have a clear understanding of the internal relationship of the network, and take targeted strategies to accelerate or inhibit the spread of information.

Our method avoids artificial subjectivity to a certain extent, because the weight of the link is measured according to the real forwarding records. We try to determine the key nodes according to the internal characteristics and objective laws of the message propagation process. The positive role of weak relationship cannot be ignored.

We may not be able to accurately describe the internal structure and interpersonal relationships of online social networks, but we can speed up or predict information dissemination by identifying the critical paths of the whole network. I hope this paper can enlighten this direction.

References

[1] Marwick A E, Boyd D . I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. 2011, New Media & Society, 13(1) 114-133.

[2] Leskovec J, McGlohon M, Faloutsos C, et al. Patterns of Cascading Behavior in Large Blog Graphs. 2007 In: Proceedings of the 7th SIAM International Conference on Data Mining, Minneapolis, Minnesota, USA. New York: ACM Press, 551-556.

[3] Gruhl D, Guha R, Liben-Nowell D, et al. Information Diffusion Through Blogspace. 2004 Proceedings of the 13th International Conference on World Wide Web. New York: ACM Press, 491-501.

[4] Zanette D H. Dynamics of rumor propagation on small-world networks. 2002 Phys Rev E, 65(4) 041908.

[5] Guille A, Hadic H. A predictive model for the temporal dynamics of information diffusion in online social networks. 2012 Proceedings of the 21st international conference on World Wide Web. ACM, 1145-1152.

[6] Lü L, Chen D B, Zhou T. The small world yields the most effective information spreading. New J Phys, 13(12): 123005.

[7] Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. 2010 ICWSM, 10: 355-358.

[8] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. System 2-10 Science (HICSS), 2010 43rd Hawaii International Conference on System Sciences.IEEE, 1-10.

[9] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?. 2010 Proceedings of the 19th international conference on World wide web.ACM, 591-600.

[10] Kahanda I, Neville J. Using Transactional Information to Predict Link Strength in Online Social Networks.2009 ICWSM, 9 74-81.