The Evaluation of Online Social Network's Nodes Influence Based on User's Attribute and Behavior

Xiushuang Yi, Yeting Han, and Xingwei Wang

College of Information Science and Engineering Northeastern University Shenyang, China
xsyi@mail.neu.edu.cn

Abstract. Objective and accurate assessment of each node influence is a vital issue to research social networks. Many algorithms have been developed, but most of them use of single metric, which is incomplete and limited to evaluate node influence. In this paper, we propose a method of evaluating node influence based on user's attribute and behavior. We study the quantification of nodes influence. The thought of PageRank is used to explore the effect of behavior. Then the method proposed is applied to Sina micro-blog. Experiment results show that method has a good and reasonable value.

Keywords: social networks, node influence, PageRank, node attribute.

1 Introduction

In last ten years, the numbers of online social networks like Facebook, MySpace and Twitter gained considerable popularity and grown at an unprecedented rate. The emergence of online social networks has brought great convenience to people's life. One of the main applications of online social networks can help manufacturers to promote their products, and to utilize the lowest investment cost to achieve the maximum effectiveness of marketing. Therefore, online social networks have also attracted many scholars to study it. A part of scholars focus on the influence of online social networks nodes, more and more interests have been made in obtaining information from social networking websites for analyzing people behaviors. These researches are focusing on identifying the influential social network users, so it can help to increase the marketing efficiency, and also can be utilized to gather opinions and information on particular topics as well as to predict the trends [1-3].

Micro-blog is a typical online social network, it not only has the social network characteristics but also has the media characteristics, so it can be analyzed from social networks and news dissemination, how to find influential micro-blog is a basic problem in the research and application of microblog.A lot of micro-blog application use "micro-blog numbers", "number of fans", "attention", "forward", as the ranking basis. These indicators can only measure the nodes from one aspect. These indexes not only can't help the user to find the influence of micro-blog quickly, but also can not truly reflect the actual influence of micro-blog users in the network. In order to overcome these shortcomings, we try to present a new method to evaluate node influence.

J. Su et al. (Eds.): ICoC 2013, CCIS 401, pp. 9–20, 2013.

In this paper we take the micro-blog for example, put forward the method of evaluating node's influence of online social networks. This paper is organized as follows: Section 2 we give an overview of related work. We introduce the basic idea of our method in Section 3. Section 4 we introduce the method of this paper in a great detail. The analysis of experiment in Section 5. We draw conclusions of this paper in Section 6.

2 Related Work

In earlier research, the people according to the method of system science use indexes like node degree, betweenness, closeness, information, eigenvector, the network diameter and so on to measure the influence. In some recent studies of influence people gradually combine the methods of social network analysis and methods of Internet search [4].

There are a lot of study of micro-blog mainly focus on the Twitter. Efforts have been made to evaluating influence of online social network [5-19]. Leila [5] investigates the power of retweet mechanism and findings suggest that relations of "friendship" at Twitter are important but not enough. Sun [6] proposes a graph model to represent the relationships between online posts of one topic, in order to identify the influential users.Jianshu Weng [7] proposed TwitterRank which measures the influence taking both the topical similarity between users and the link structure into account.Meeyoung analyzed Propagation characteristics of Twitter, micro-blog forwarding and uses three parameters, by the study of a large number of Twitter data [8]. So they found effects of the user in the topic in the process of communication. Pal performed an extensive study about Twitter follower-following topology analysis [18].

Wu [10] utilize power multiplication iterative to calculate Markov matrix, by optimizing and improving the PageRank algorithm. Yang [11] Starting from the two angles of active users and blog quality, constructed the evaluation index of the blogger influence, introduced the blogger communication ability factor, using the idea of PageRank algorithm to design a new influence ranking algorithm to evaluate the blogger influence.Guo [12] proposed the quantitative definition of user information dissemination scope, and gives the method for calculating the influence.

3 The Basic Ideas of Algorithm

As we know, every micro-blog users in the network corresponding to individual or unit of reality. User can enhance his own prestige by publishing micro-blog, forwarding and commenting of others, concern for others. The attribute in the micro-blog included two parts e.g. user attributes and micro-blog properties. The user itself includes the user ID, user type, attention number, number of fans , number of micro-blog, number of mentions, and micro-blog attributes including number of micro-blog, publish time, the forwarding numbers, numbers of comments. Micro-blog network and online community network, the user can according to their own preferences selectively use "forward", "collection", "comment" on a piece of information or micro-blog do corresponding operations.

Node's attribute is the basic characteristic of node. If the user measure the node's influence only by itself or micro-blog attribute to measure the node's influence is

relatively one-sided, they should also be take the direct effect on node influence which is given by behavior between nodes in account. The behavior of forward and comment can change the size of the impact of a user. Therefore, when measure of influence of nodes we should fully consider the interaction between the nodes that play an important role for influence.

We proposed model of influence rank (model of IR).The method in this paper proposed by combining the interactive behavior between nodes and node's attributes. Firstly we need to give a node attribute's quantification. Node attribute has many factors; these factors will influence effect of node. We should select some main influence factors, and then we use analysis hierarchy process to give the weight of every factor of influence, using the weighted to calculate the quantification of node attribute value. Secondly we introduce the thought of PageRank [13] algorithm to study the node interaction behavior. The PageRank algorithm is based on the assumption: the webpage is more important when it link to more webpage; webpage is more important when it linked to the more important webpage. Similarly, for online social networks, when the user is more commented by others, its influence will be greater; when the user is forward comments by more important users; its influence will be greater. Finally the user attributes and user behavior as the node influence factor synthesis node's influence in online social networks.

4 The Evaluation of Node's Influence

4.1 Measuring Node Attributes

There are lots of node's attributes in online social network; we select some attributes which have more obvious role: user type, numbers of fans, numbers of forwarded, numbers of attention, numbers of micro-blog, and numbers of comments. The numbers of fans, numbers of forwarding, and the numbers of comments can reflect the influence of nodes from different aspects. The analytic hierarchy process to solve the weight problem of each influence factor, and then use the weighted and calculate the quantization node attribute value. Calculation steps as follows:

Step1: Construction of index matrix **X** and normalization, so we get new matrix **A** as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

We can use the standard 0-1 transform to normalize:

$$a_{ij} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_{ij}} \tag{1}$$

Where $x_j^{\max}$ the maximum value in the j column is, $x_j^{\min}$ is the minimum value in the j column.

Step2: Using analysis hierarchy process to calculate the index weight.

(a) Constructing a judgment matrix **B**, we use Table I to determine the value of **B**.

Table 1. The Reference Values of Elements in Matrix **B**

Relative importance	Definition	Meaning
1	Equally importance	Two attributes are equally importance
3	Slightly importance	One attribute is more slightly importance
5	Considerable importance	One is considerable importance
7	Obvious importance	One is more obvious than another
9	Absolutely importance	One is more absolute importance
2、4、6、8	Compromise	Compromise of two grades

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & \cdots & b_{mm} \end{bmatrix}$$

The elements b_{ij} are scale of b_i relative to b_j in matrix **B**.

(b) To determine the weight of different indexes and check consistency:

(i) Multiplication element of B in a row:

$$b_i = \prod_{j=1}^{m} b_{ij} \quad j = 1, 2, \cdots, m \tag{2}$$

(ii) Calculating product to the m[th] roots:

$$c_i = \sqrt[m]{b_i} \quad i = 1, 2, \cdots, m \tag{3}$$

(iii) Weight calculation:

$$w_i = \frac{c_i}{\sum_{j=1}^{m} c_j} i = 1, 2, \cdots, m \tag{4}$$

(iv) Checking consistency: we compute the latent root of **B** and calculate coincidence index denoted by *CI*.

$$CI = \frac{\lambda_{\max} - m}{m-1} \tag{5}$$

If $CI < 0.1$,we can accept the weight of different indexes. Otherwise, we need to recalculate the **Step2**.

Step3: Calculating AR according to the weight of each index w_j and the index value of each node a_{ij} :

$$AR(i) = \sum_{j=1}^{m} a_{ij} w_j \quad i = 1, 2, \cdots, n \tag{6}$$

4.2 Behavior Measurement

In this process, we use PageRank algorithm to solve the problem of interaction between nodes.PageRank is one of the core technology of Google, its basic idea is to determine the importance of using webpage hyperlink structure webpage, The PageRank generates represent each page importance value becomes the PR value, a page's PR value not only depends on the number of connected to the page, but also be influenced by the quality and importance of the page, but the page's PR value will be evenly distributed to its chain of the page[14].The calculating formula of PageRank as follows:

$$PR(p_i) = d + (1-d) \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{7}$$

Where $L(P_j)$ is number of P_j links to other page, where $M(P_i)$ is numbers of chain into P_i , d is damping coefficient.

Then extract the relationship of forwarded or comments in online social network, a directed graph of forward comments relationship could be constructed. In the following simple diagram, the direction of the arrow represents the object forwarding comments. Because of the large scale network, the analysis is difficult; we can start from a node to consider. Observation of the following figure, for example the user a forward or comment of user e、 f and g. Of course, there are also have other users will forward information of user a. The user e and g may also go forward or comment on other user, this part of the forward or comment leave out in Fig.1.We consider only the relationship between local forward or comments, and then further promotion.

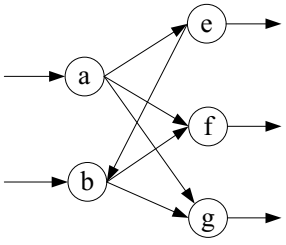

Fig. 1. Schematic diagram of forwarding

For example, user a and user b have forwarded f, and the user a also forwarded user e. N_{af} is the number of a forwarded f, N_{bf} is the number of b forwarded f, N_{ae} is the number of a forwarded e. $BR(a)$ and $BR(b)$ are important degree. We can make conclusion as following (I) if the number of a forwarded f is more i.e. $N_{af} > N_{bf}$. It is suggest that user a make a larger contribution to f. (II) if user a is more important than b i.e. $BR(a) > BR(b)$ .The user a will make a larger contribution to f. (III) The user a forwarded e and f, but if $N_{af} > N_{ae}$, it suggest that user a make a larger contribution to f.

These three conclusions obtained above are consistent with the idea of PageRank algorithm; the interactive behavior can change the node influence. Forwarding can be as the voting behavior, the more number of forwarding equivalent to more votes, "the greater support to other users", such as the above conclusions (I) (III); if it is an important vote, the "support" will be greater, likes the above conclusion (II). Unlike the original PageRank algorithm, the node "contribution" in this paper is not the average distribution, in the conclusion (III) have illustrated this point. Distribution of the "contribution" is related to the number of forward or comments; the more forward the more "contribution". For the general case, If the relationship of forward or comment between nodes v and u is existed, the node v forward node u, then use the following formula to express the v to u "contribution" ratio:

$$p_{vu} = \frac{N_{vu}}{\sum_{x \in O(v)} N_{vx}} \tag{8}$$

The N_{vu} stand for number of *v* forward u, $O(v)$ is the set of *v* forwards, N_{vx} is *v* forwards others, where p_{vu} represents the probability of v forward u.

The core idea of PageRank is to calculating the PR value of each node value according to the number of back links, uniform "flow" to all nodes. Each node's PR value is the total of "contribution" of node's PR value. We borrowed this idea. Then we revised the problems of "contribution" average distribution, and put forward the calculating formula evaluation node behavior influence:

$$BR(u) = (1-d) \times BR(u) + d \sum_{v \in I(u)} p_{vu} BR(v) \tag{9}$$

This is an iterative formula. The initial value of *BR* is related to *AR* refer to formula (6), and where p_{vu} is distribution probability, where $I(u)$ is set of pointed to *u* in directed graph. The value of d is 0.85.The formula (9) combines the characteristics of online social network and its application background.

The value of BR represents the influence of behavior; the calculation process of BR is following:

Calculation of the value of BR

Get data set G=(V,E),where V is users set, E is relationship set, initial AR, BR, N_{vu}, P_{vu}, d, ε, σ

for each user $v, u \in V$ do
  update value of N_{vu}
end for
for each user $v, u \in V$ do
  $P_{vu} \Leftarrow N_{vu} / \sum_{x \in o(N)} N_{vx}$
  update value of P_{vu}
end for
$i \Leftarrow 1$, $BR_1 \Leftarrow AR$, $\varepsilon \Leftarrow 0.0001$, $\sigma \Leftarrow 1$
while $\sigma \geq \varepsilon$ do
  $BR_{i+1} \Leftarrow (1-d) \times BR_i + d \times P \times BR_i$
  $\sigma \Leftarrow \| BR_{i+1} - BR_i \|$
  $i \Leftarrow i + 1$
end while
return BR

4.3 To Establish the Model of IR

The first two chapter studied effects on influence of nodes from aspect of node attributes and interaction behavior. According to the above algorithm, the two aspects will be integrated. We use them as nodes influence factor synthesis in online social networks influence.

After a lot of observation and analysis: node attribute is a basic condition determines the influence of nodes, if the node attribute value AR is weak, the node will not influence overall strong. At the same time, the interactive behavior between the nodes can also change the node influence; this kind of behavior is equivalent to the influence in the network spread, similar to the random walk model. The interaction behavior between nodes and the attributes of the nodes combined in accordance with the appropriate weight, a formula to calculate the influence of nodes:

$$IR(i) = \alpha AR(i) + (1-\alpha) BR(i) \ \ i = 1, 2, \cdots, n \tag{10}$$

In formula (10) α is regulatory factor, it can regulate the weight of AR and BR. This formula is a linear combination of AR and BR. When $0.5 < \alpha < 1$, it is shown that the node attribute is more importance. When $0 < \alpha < 0.5$, it shown that we more emphasis on interactive behavior between nodes to measure the influence.

5 Experiments

The paper selects the Sina micro-blog as the data source, the data obtained by Sina open API.The data set is part of Sina micro-blog user information in December 2012. It contains 60290 users. Every user includes a user ID, user type, numbers of fans, numbers of micro-blogs, numbers of attention, forwarding numbers, the number of comments.

Firstly, we construct matrix of the influence factor, and utilize analysis hierarchy process to determine the weight of each node, and then calculate node's *AR*.According to the network relationship of micro-blog, and the matrix of transition probability, we can get the node's *BR*.Finally, and we calculate the node of the *IR* value through Influence Rank model proposed in this paper.

Table 2. The Calculation Results of Top 10 Users

UserID	AR	BR	IR
1266321801	1.0000	0.4409	0.6086
1192329374	0.7045	0.3538	0.4590
1656809190	0.5305	0.4265	0.4577
1087770692	0.5508	0.3814	0.4322
1192515960	0.3304	0.3977	0.3775
1752467960	0.3249	0.3390	0.3347
1742727537	0.1000	0.4233	0.3263
1730336902	0.1331	0.4046	0.3231
1212812142	0.2137	0.3683	0.3219
1854283601	0.0601	0.4126	0.3069

Because the data set is large, we list the influence of top ten users. In the process of calculation, we assumed α is 0.3.Node's *IR* value is not only closely related with the *AR* and *BR* values, but also with the relevant to the value of α. In general, if the influence of node is greater, the *AR* and *BR* will be greater.

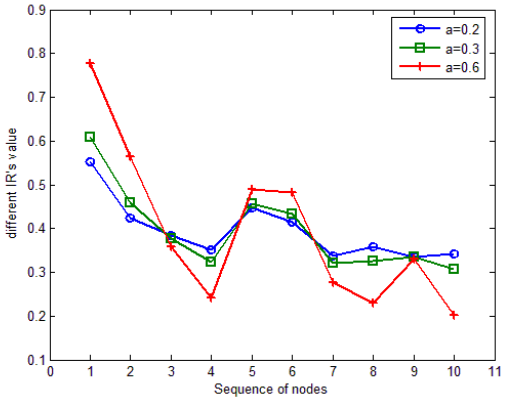

Fig. 2. Comparison of different IR

Like formula (10), when α taking different values, the results of IR are different. As is shown in Figure2 the node's IR value is closely related to regulatory factor α. The regulatory factor represents the weight of AR.As is shown in TABLEII, the 10^{th} node's AR is very small, when α is larger the IR will be smaller as shown in Figure2. Therefore, α is used to balance of AR and BR according to need.

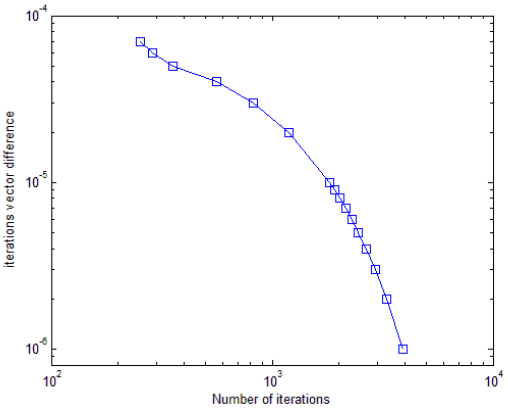

Fig. 3. Convergence analysis of BR

According to the core idea of PageRank, the calculation results of BR have nothing to do with initial value. As is shown in the Figure3, after 3928 times of computation, the difference vector is close to 0. It shows that the value of BR gradually trends a stable value. Therefore, the value of BR is convergence. According to formula (10), the value of IR also trends a stable value. So that the IR' value of each node can be calculated.

As is shown in the Figure4, we use three methods to calculate the influence of node. This shows that our method is closely related with the number of fans and the number of micro blogs. Observed from the curve of fans, fans of the 3rd node are larger than the 4th node. But 3rd node's IR is less than 4th node's IR, from another perspective, the 4th node's interactive behavior is greater than 3rd node's, this shows that the 4th node is involved in more social behavior which can bring greater influence. Observed from the curve of comments, the 5th node's comments is less than 6th node's but 5th node's *IR* is about as same as 6th node's *IR*. It is shows that 5th node's AR is larger. The experimental results have shown that IR is related to user's attribute and user's behavior.

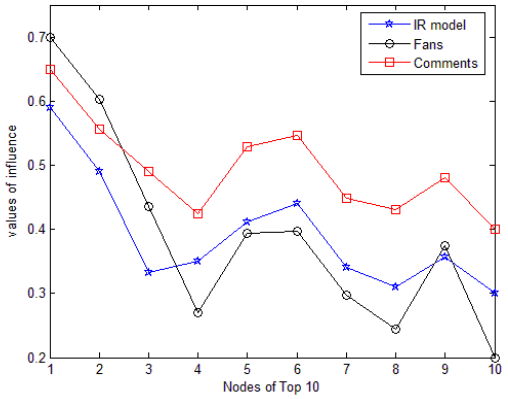

Fig. 4. The relation between IR model and Fans as well as that between IR model and Comments

We use the other two methods to evaluate the influence of node: UserRank [9] and TURank [15].These methods are based on PageRank algorithm. The core idea of UserRank is that the numbers of friends is an important index of influence. The TURank could to reveal the relationship between user and information and obtain a more accurate result of ranking. The experimental results are as follows.

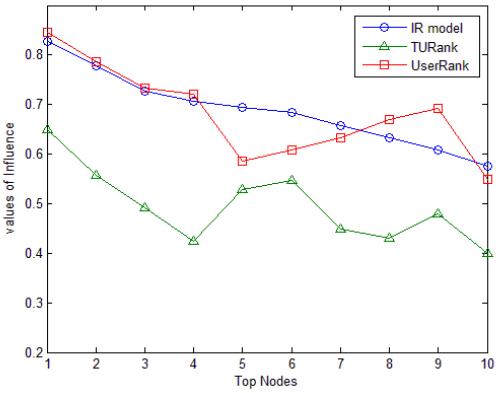

Fig. 5. Comparision of three methods

We selected the top ten users as shown in TABLEII. The value of IR gradually decreased from node 1 to node 10. Observed from the curve of UserRank, the first four nodes' influence is decreasing, but the influence of subsequent node is increasing, the influence of node 9 is very large. Why trends of the two curve is not consistent? The most important reason is that the two methods considering the problem from different angles. The method of UserRank focuses more on the number of friends. For example, the experiment shows that node 9 has 876 friends so that its influence is larger. Because of the existence of zombie fans, the node has a lot of friends and its influence unnecessarily large. Compared with IR model, UserRank is not considered comprehensively, the result of UserRank is not very accurate.

Observed from the curve of TURank, the first four nodes' influence is decreasing, the influence of subsequent node vary irregularly. The trends of the IR model and TURank are not consistent. The method of TURank pays more attention to interactive behavior and information of user own. As is shown in the Figure5, the node 6 has a large value of TURank; because of the node has more interactive behavior. The method of TURank ignored the number of fans and its initial value should be set artificially. Therefore, the results of two methods are not completely consistent. Compared with TURank, the result of IR model meets the actual better.

From the results of these experiment can be seen, the evaluation of nodes influence based on user's attribute and behavior considering many factors, and the result is more realistic. This method has certain applicability.

6 Conclusion

This paper proposes the model of IR; it can be comprehensive consideration attributes and interactive behavior. Firstly, we calculate the node attribute value AR, and then

calculate the node's BR according to the principle of PageRank.The model avoids the defects caused by using of single factor to evaluate node's influence. Experiments show that node's IR has no direct relationship with fans and numbers of micro blog. Theory analysis and example of real network experiment show that the new proposed method can effectively evaluate the node influence of the online social network.

Acknowledgment. This work has been supported by the National Natural Science Foundation of China under Grant No.61070162 and No.60903159; the National Science Technology support Project of China under Grant No.2008BAH37B05;the National High Technology Research and Development Project of China under Grant No.2007AA041201;the Project of Fundamental Research Funds for Central-affiliated University of China under Grant No.N110216001.

References

1. Xu, Z., Lu, R., Xiang, L.: Discovering User Interest on Twitter with a Modified Author-Topic Model, pp. 422–429 (2011)
2. Ghosh, S., Viswanath, B., Kooti, F.: Understanding and combating link farming in the twitter social network. In: Proceeding of the 21st International Conference on World Wide Web, pp. 61–70 (2012)
3. Benevenuto, F., Rodrigues, T., Cha, M.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, pp. 49–62 (2009)
4. Sun, R., Luo, W.: Review on evaluation of node importance in public opinion. Application Research of Computer 29(10), 3606–3608 (2012)
5. Weitzel, L., Quaresma, P.: Measuring node importance on Twitter microblogging. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, vol. (11) (2012)
6. Sun, B., Ng, V.T.: Identifying Influential Users by Their Postings in Social Networks. In: Proceeding of the 3rd International Workshop on Modeling Social Media, pp. 1–8 (2012)
7. Weng, J., Lim, E.-P.: TwitterRank:Finding Topic sensitive Influential Twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270 (2012)
8. Meeyoung, C.: Measuring user influence in twitter: The million follower fallacy. In: Proceedings of International Conference on Weblogs and Media (2010)
9. Jun, L., Zhen, C., Jiwei, H.: Micro-blog Impact Evaluation. Information Network Security (3), 10–13 (2012)
10. Jialin, W., Yongji, T.: The optimization and improvement of PageRank algorithm. Computer Engineering and Applications 45(16), 56–59 (2009)
11. Changchun, Y., Kefei, Y., Shiren, Y.: New assessment method on influence of bloggers in community of Chinese microblog. Computer Engineering and Applications 48(25), 229–233 (2012)
12. Hao, G., Yuliang, L., Yu, W.: Measuring user influence of a microblog based on information diffusion. Journal of Shandong University (Natural Science) 47(5), 78–83 (2012)
13. Xiaofei, C., Yitong, W., Xiaojun, F.: An Improvement of PageRank Algorithm Based on Page Quality. Journal of Computer Research and Development 46(suppl.), 381–387 (2009)

14. Shenjun, Z., Xiongkai, S.: An Improved N-PageRank Algorithm which Considers the User Behavior. Computer Technology and Development 21(8), 137–140 (2011)
15. Yamaguchi, Y., Takahashi, T., Amagasa, T., Kitagawa, H.: TURank:Twitter User Ranking Based on User-Tweet Graph Analysis. In: Chen, L., Triantafillou, P., Suel, T. (eds.) WISE 2010. LNCS, vol. 6488, pp. 240–253. Springer, Heidelberg (2010)
16. Ye, S., Wu, S.F.: Measuring Message Propagation and social Influence on Twitter.com. In: Bolc, L., Makowski, M., Wierzbicki, A. (eds.) SocInfo 2010. LNCS, vol. 6430, pp. 216–231. Springer, Heidelberg (2010)
17. Sun, B., Ng, V.T.: Lifespan and Popularity Measurement of Online Content on Social Networks. In: Social Computing Workshop of IEEE ISI Conference, pp. 379–383 (2011)
18. Ilyas, M.U., Radha, H.: A KLT-inspired Node Centrality for Identifying Influential Neighborhoods in Graphs. In: Conference on Information Sciences and Systems, pp. 1–7 (2010)
19. Pal, A., Counts, S.: Indentifying topical authorities in microblogs. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 45–54 (2011)