International Journal of Database Theory and Application
Vol.9, No.1 (2016), pp.201-208
http://dx.doi.org/10.14257/ijdta.2016.9.1.17

# Topical Influence Analysis Algorithm based on Information Propagation in Microblogs

LinTao Lv and QinQin Yuan

*Dept. of Electronical and Information Engineering, Xijing University, china*
*lvlintaoxijing@sina.com, jeanxun7910@163.com*

## *Abstract*

*With the development and popularization of the Internet, the different microblog topics are disscussed everyday, so the microblog can product a large number of various topics,which can reflect the influence of different users in a given topic. In the microblog topics,the key users of microblog topics are found by discovering the influential sensitive topics and calculating the influence value of the user,which are the focus of attention in the fields of microblog public opinion supervision and safety management.*

*In order to accurately measure the influencers in the given topic and to calculate the user influence value, the thesis proposes the method of constructing the propagation network which is based on attention and forward relationship between users, and then proposes TDN-If algorithm by using PageRank algorithm. When we calculate the transition probability in propagation network by using the TDN-If algorithm, the information propagation is considered to measure influencers. This method can resolve the defects of discovering the user's influence by only using followers of this single indicator in the current microblog topics. The experimental results show that the TDN-If algorithm has important theoretical and practical value, which is better than TwitterRank algorithm and other influential individuals found algorithm.*

*Thus, the method proposed in this paper can not only effectively solve the problem about discovering and persuading the key users in the sensitive topics who have influences and have the unique insights on the significant events, for example, which can provide the strong guarantee for the governments in the fight against terrorism, but also provide the important theory and method for the complex network community discovery, microblog public opinion supervision , microblog safety management and so on.*

***Keywords:*** *information propagation, attention, forward, PageRank, influencer*

## 1. Introduction

With the development and popularization of the internet, every aspect of people's lives has been inseparable from the network media. Many people may participate in different microblogs topics every day, some of which key user guide on topics played a key role. Finding the key user in the sensitive topic, and to ease, which has became the focus of public opinion regulatory authorities.

The influencers of microblogs topic are able to provide the valuable information and have unique viewpoints for important events. These users are key users of the hot topics, which can provide support for the analysis of network public opinion by analyzing the characteristics of those users. In the early stages of Twitter studies, the number of followers is used to directly measure the user influence [1-2]. The method of measurement is simple, but the form is too single, which has certain deviation comparing to user actual influence. Some scholars have found the influential individuals based on the microblog network structure using PageRank [3]. Typically Topic-sensitive PageRank [4], which is mainly different from PageRank is that the probability of surfers random

ISSN: 2005-4270 IJDTA

jumping to different nodes has theme selection, so surfers can change randomly the subjects of interested web page , it can improve the accuracy of the algorithm to some degree. The literature [5] which is considering comprehensively the similarity and network structure of the topics attention by users, and extends the PageRank algorithm, the TwitterRank algorithm that is a new influencer ranking method is proposed based on Twitter.

From the above research, it can be seen that the information propagation features can reflect more visually the influencer. But the analysis and research of the influencer is currently less. So the thesis tries to find the key users in the topics in information propagation.

## 2. Network Construction in the Information Propagation

The topic propagation is essentially similar to the overall performance of the microblog information propagation, the effect of information propagation reflects directly the influencer. So the research of the information propagation can analyze more accurately the user influence. The thesis examines the influencer in the information propagation by considering comprehensively behavior among users and the situation of information forwarding.

### 2.1 Microblog Information Propagation Mechanism

In Twitter service, the mechanism of microblog information propagation has mainly two type: information pushing behavior and user forwarding behavior. If the user $u_0$ releases one microblog message Mes to the user collection $\overline{U} = \{u_1, u_2, \cdots\}$, saying that the message Mes is from the user $u_0$ to $\overline{U}$, and $\overline{U}$ is the fans collection of user $u_0$. When the user $u_0$ publishes one microblog message, then the microblog will be automatically pushed to all fans pages of the user $u_0$ , which is the mechanism of microblog message pushing. If the user ($u_1 \in \overline{U}$) receives the message Mes and forwards it, at the same time, another user ($v \notin \overline{U}$) receives message Mes from user $u_1$, so the message Mes passes through two-stage forwarding, which is expressed as $u_0 \rightarrow u_1 \rightarrow v$. This is forwarding behavior of users, which reflects the radiation degree of influencer.

The selection of the network structure affects directly the mechanism of the microblog information propagation, in order to describe more accurately the process of the information propagation, This paper about the mechanism of the microblog information propagation has two ways of obtaining the structure of the network, the one that is obtaining attention network based on the social relationship of users; the other that is obtaining forward network based on user behavior.

### 2.2 Attention Network Rebuilding

The rebuilding of attention network is described as follows: in a topic, we can firstly obtain the user uid involving the topic according to the microblog information about the topic (include “//@username” in microblog), and obtain attention list of the user according to the user uid, the users of the list are users of the topics collection, then the attention network will be obtained according to the attention relationship.

### 2.3 Forward Network Rebuilding

The rebuilding of forward network is described as follows:

(1) if the microblog A contains "//@ microblog B", then establish an edge between node A and node B, that is, B->A;

(2) if the microblog contains multiple "//@", indicating nested forwarding, the directed edges is built successively in accordance with the order of the "//@";

(3) The weights of B->A is the number of microblog B forwarding to microblog A in a topic.

### 2.4 Information Propagation Network Rebuilding

The Study is found that the user behavior is not well described by the network based on attention relationship. It is proposed in the thesis that forward network is mapped to the attention network, and the influencer is quantitatively analyzed in topic propagation network. Thus, according to the forward network and the attention network, the rebuilding of the topic propagation network is described as follows:

input: the attention network$G_L(V_L, E_L)$, the forward network$G_F(V_F, E_F)$ ;

Output：the topic propagation network $G_T(V_T, E_T)$;

Step 1: input $G_L(V_L, E_L)$and$G_F(V_F, E_F)$ ;

Step 2: initialize$G_T(V_T, E_T)$, set $G_T(V_T, E_T) = G_L(V_L, E_L)$;

Step 3: an initial value n is given for the element of $E_T$;

Step 4: take any node $u \in V_F$

Step 5: take any element $e^m_{F(u \to v)}$ of the collection$OE_F(u)$;

Step 6: if exists $e^n_{T(u \to v)} \in E_T$,then set n=m+n; if $u, v \in V_T$ but $e^n_{T(u \to v)} \notin E_T$ ,$e^n_{T(u \to v)}$is taken into $E_T$,and set n=m; if $u, v \notin V_T$,u and v are taken into $V_T$,and $e^n_{T(u \to v)}$is taken into $E_T$.then goto step 5, Until all the elements are traversed completely;// Breadth-first traversal

Step 7：repeat setp4 to step6, until all the elements of $V_F$ are traversed completely;// Breadth-first traversal

Step 8：output $G_T(V_T, E_T)$.

In the Step3, the initial value of each element in $E_T$ is not fixed, for any $e_{T(u \to v)} \in E_T$, the user u sends information to the page of another user v, the microblog will be forwarded by user v in a certain probability, the fans number of user v reflects the user's contribution to the propagation of the microblog, therefore, take$n = \sqrt[3]{v_{FanNumber}}$ ;

In the Step 5, $OE_F(u)$is a directed edges collection which is chained by the node u; $e^m_{F(u \to v)}$is the element of $OE_F(u)$, m is the Weight on the edge, $e^n_{T(u \to v)}$ is Similar, the more the value of N, the more the fans, the greater the contribution to the information propagation, the greater the influencer.

## 3. Calculation of Transition Probability in Information Propagation Network

Definition 1: According to the topic propagation network structure, drawing on PageRank [6] thought, the algorithm for the influencer in the network is established, and the algorithm is defined as influencer discovery algorithm for information propagation network TDN-If (Topic Diffusion Network- Influence).

Definition 2: In the TDN-If algorithm, a transfer matrix (matrix transition) M is required to describe the transfer probability between nodes. If the total number of nodes in the topic propagation network is n, then the matrix M is a square matrix of n×n, $p_{ij}$ expresses the probability that the node j is transferred to the node i.

Definition 3: Influencer can be seen as a resource, which selects randomly a node into the network, then walks randomly along the direction of information flow between nodes, repeating the process until the probability of the influencer among the nodes tends to be stable. The position of the influencer inside the topic propagation network is described by position vector V. V is a n-dimensional column vector, the component j represents the probability of the influencer of the node j. For the influencer randomly selecting the node into the network, so the initial position vector is $v_0=(1/n,1/n,\cdots,1/n)^T$ , assuming that the position vector of influencer after T times transfer in the topic propagation network is $v_t$ ,its iterative formula is:

$$v_{t+1}=bM\cdot v_t+(1-b)\frac{e}{n} \quad (1)$$

The b is a fixed value, which expresses the probability of random jumps, ranging generally between 0.8 to 0.9, which selects 0.85 in the thesis. E is a random jump vector which is a unit of n-dimensional column vector. $bM\cdot v_t$ represents that the random surfer choose a link from the current page to continue to browse in the probability of b. $(1-b)\frac{e}{n}$ represents that the surf jumps randomly in the probability of 1-b.

For $v_{t+1}$ which is the i-th component in $s_{t+1}^{i}$, its iterative formula is:

$$s_{t+1}^{i}=\frac{1-b}{n}+b*\sum_{j=1}^{n}\left(p_{i\,j}\cdot s_t^j\right) \quad (2)$$

Definition 4: In the topic propagation network, the weights $w_r$ reflects the frequent degree of information flow between nodes. The bigger the weight is, the closer the interaction between nodes is, the greater the interest among nodes is, the greater the influence it will be, so the probability of the influencer transiting along the edge is the greater. Therefore, the transition probability is defined as the formula (3):

$$p_{i\,j}=\frac{w_{j\to i}}{\sum_{k\in O(j)}w_{j\to k}} \quad (3)$$

The $w_{j\to i}$ is the weight of $e_{j\to i}^{w}$, which is the value n in the topic propagation network, $O(j)$ is the node collection of node j chained point.

With the formula (3) substituting into formula (2), and the iterative formula of the influence node is as follows:

$$PR_{t+1}(i)=\frac{1-b}{n}+ \\ b*\sum_{j=1}^{n}\left(\frac{w_{j\to i}}{\sum_{k\in O(j)}w_{j\to k}}*PR_t(j)\right) \quad (4)$$

The $PR_t(i)$ is the influencer value of the node i after t+1 times iteration, $PR_t(j)$ is the influencer value of the node j after t times iteration. The PR value of the node is constantly iterated by the formula (4) until it begins to converge, the convergence value reflects the influencer of the nodes.

## 4 Experimental Results and Analysis

### 4.1 Experimental Data

(1) Experimental platform: The hardware environment of this experiment is composed of Windows XP SP3、Inter(R) Core(TM)2 Duo CPU E7400 @2.80GHz and 2GB memory.

(2) Experimental data: Microblog related information is obtained with the help of the corresponding API interface which are provided by Sina Weibo. With the help of the corresponding API interface, ten kinds of hot topic microblog data in Sina Weibo are

obtained, including user related properties, user microblog information, user behavior, *etc.*

### 4.2 Evaluation Indicator

Definition 5: Because the actual influencer of the microblog user is difficult to analyze precisely [7], accuracy and recall of TDN-If algorithm is verified crossly by a variety of algorithms in the thesis.

Definition 6: The comparison between several influencer mining algorithms and the algorithm presented in the thesis is as follows:

(1)RepostRank algorithm, the influencer in the forward network which is formed by relying on forward relationships, is discovered by PageRank algorithm.

(2) TwitterRank, *et al.,* [8] proposed Weng algorithm, when computing the transition probabilities, the algorithm considers the topic similarity and user activity among users , but the jump probability of the surfer remains fixed, so the algorithm is not effective in dealing with the "acquisition trap";

(3) In-degree algorithm, which relies on the number of the user followers to measure influencer, this algorithm is currently used in Twittter and other third party services, such as twitterholic.com, wefollow.com;

(4) TweetNum algorithm, which relies on the number of posts to measure user influencer;

(5) TDN-If algorithm, which is influencer discovery algorithm.

In the experiment, the method of cross-validation is adopted, that is, a variety of (N) algorithms are considered the correct results as the reference results. For example, given four kinds of algorithms : A, B,C,D, obtained Top K influencer collections : $I_A$, $I_B$, $I_C$ , $I_D$, it is assumed that the two algorithms are considered the correct results as the reference results, the influencer reference standard collection is defined as shown in the formula (5):

$$I_2 = (I_A \cap I_B) \cup (I_A \cap I_C) \cup (I_A \cap I_D) \\ \cup (I_B \cap I_C) \cup (I_B \cap I_D) \cup (I_C \cap I_D) \qquad (5)$$

The P( precious) Can reflect the real situation of the influencer in the micro-blog, that is, the accuracy of the discovery for the influencer discovery by A, algorithm, which is defined as shown in the formula (6):

$$P_A = \frac{|I_A \cap I_2|}{|I_A|} \qquad (6)$$

The R(recall) can reflect the degree of the discovery for the influencer in the micro-blog. that is, the recall of the influencer discovery in the algorithm A, which is defined as shown in the formula (7):

$$R_A = \frac{|I_A \cap I_2|}{|I_2|} \qquad (7)$$

The F1 can consider the accuracy and recall, which can reflect the overall accuracy of the algorithm and the degree of the recall. The F1 of the influencer discovery in the algorithm A, which is defined as shown in the formula (8):

$$F1 = \frac{(\alpha^2 + 1)P * R}{\alpha^2 (P + R)} \qquad (8)$$

F1 is the most common value, when a=1, which is selected in this paper.

## 4.3 Experimental Result

### 4.3.1 Accuracy Verification

According to the above five kinds of algorithms, the Top 10, Top 20, Top 50, Top 100 influencer in each topic are acquired. For the N=2, 3, 4, selecting Top 10, Top 20, Top 50, Top 100 in each topic ,the average accuracy of influencer mining algorithm are shown in Figure 1, X axis represents 10 kinds of different topics.

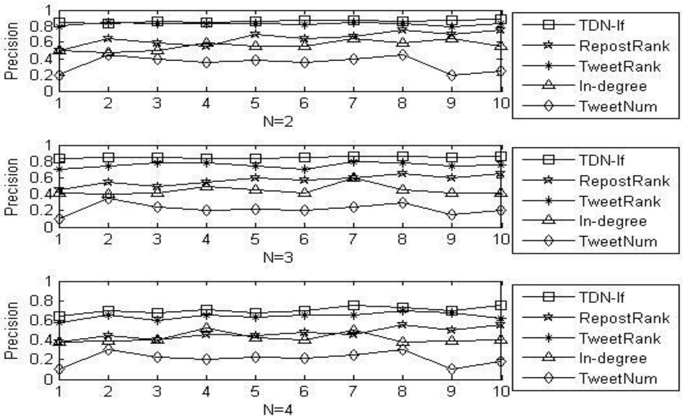

**Figure 1. The Average Accuracy of Five Kinds of Algorithms in All Rankings**

Experimental results show that, when n = 2, 3, 4, the accuracy of TDN-If algorithm is relatively high in most of the topics, TweetNum algorithm has the lowest accuracy. This is because that TweetNum only considers the publishing number of Twitter users, without considering the quality of microblogs, therefore, the accuracy of discovering influencer is not high. And TDN-If is according to the forwarding relationships of information and attention relationships of user to find influential individuals, not only considers the microblog quality (the forwarding number of microblog daily) ,but also considers the influence of fans, so influencers discovered are more accurate and closer to the actual situation .

Experimental results also show that, in each topic, with the increase of N, which is the number of cross validation reference standard, the accurate is decreased. This is due to the increase of the standard reference number N, the elements of the influence reference standard collection $I_N$ become smaller, then the elements of the collection $I_A \cap I_N$ is decreased, so that the accuracy of the algorithm is reduced; The experimental results also show that when N=3, the accuracy of five kinds of algorithms is obvious, and the experimental results are the best. If N is too low, the elements of the influence standard reference collection $I_N$ will be more, the accuracy may be higher than the actual situation; If N is too high, the elements of the influence standard reference collection $I_N$ will be less, then lead to be basically same between each algorithm and the conjunctional elements of the standard reference collection $I_N$, so the accuracy is not obvious.

### 4.3.2 Recall Verification

Also the Top 10, Top 20, Top 50, Top 100 influencer in each topic are acquired. For the N=2, 3, 4, selecting Top 10, Top 20, Top 50, Top 100 in each topic ,the average recall of influencer mining algorithm are shown in Figure 2.

Experimental results show that, when N=2, 3, 4, TDN-If algorithm has a high recall in most of the topics. Experimental results show that, with the increase of N, recall is increasing. This is due to the increase of N, the element of the influence reference standard collection $I_N$ is reduced, that is, the number of the influencer referenced is reduced, then the more influencers will be found, so the recall will be higher. Experimental results also show that, when N=3, the distinction for the recall of each algorithm is greater, and the experimental results are the best. If N is too small (N=2), the elements of the collection $I_N$ become more and more, which makes the elements of $I_A \cap I_N$ basically consistent, so that the recall rate is not large.

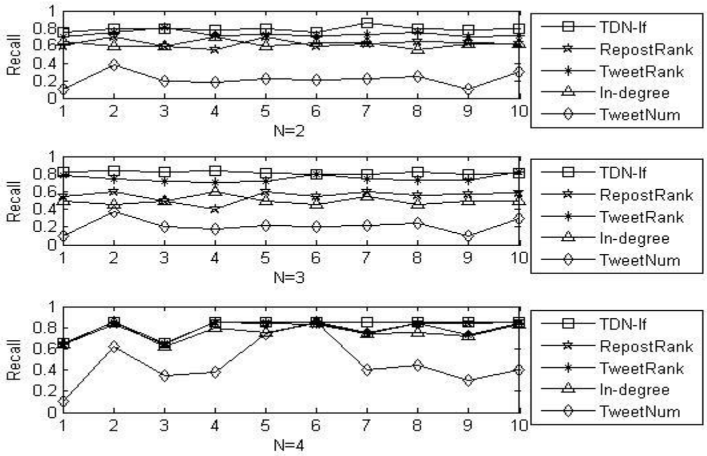

**Figure 2. The Average Recall of Algorithms in All Rankings**

### 4.3.3 F1 Verification

Also the Top 10, Top 20, Top 50, Top 100 influencer in each topic are acquired. For the N=2, 3, 4, selecting Top 10, Top 20, Top 50, Top 100 in each topic ,the average F1 of influencer mining algorithm are shown in Figure 3.

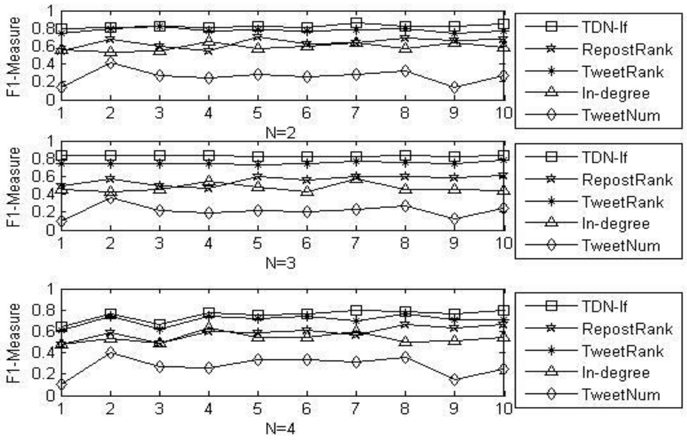

**Figure 3. The Average F1 of Five Kinds of Algorithms in All Rankings**

Experimental results show that, when N=2, 3, 4, the F1 value of TDN-If algorithm in most of the topics are higher, which reflects the overall performance of the algorithm. From the graph, it can be seen that when N=3, the overall performance of the algorithm is obvious, and the experimental results are the best.

## 5. Conclusion

The user's attention behavior and forwarding behavior are the main way of the microblogs information propagation. In this paper, the forwarding network is mapped to the attention network, which is good to solve the problem that user social relations can not accurately describe the way of the information propagation. The TDN-If algorithm is proposed to calculate the user influence based on the PageRank algorithm, which has higher accuracy and recall comparing with the RepostRank, In-degree, TweetRank and TweetNum algorithm. In this paper, the accuracy and the effectiveness of the algorithm are considered, and the performance optimization of the algorithm will be further discussed and studied by the application of parallel computing technology in the processing of microblogs big data.

## Acknowledgement

This research was sponsored by the National Natural Science Foundation of China (61273271)，Teaching reform project of xijing university of china(JGGH1512)，scientific research project of xijing university of china(XJ150122).

## References

[1] H. Liangjie and B. D. Davison, “A classification-based approach to question answering In discussion boards”, Proceedings-32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, United states, **(2009)** July19-23, pp. 171-178.

[2] B. Krishnamurthy, P. Gill and M. Arlitt, “A few chirps about twitter”, Proceedings of the ACM SIGCOMM 2008 Conference on Computer Communications-1st Workshop on Online Social Networks, Seattle, United states, **(2008)** August 17-22, pp. 19-24.

[3] B. De La Ossa, J. A. Gil, J. Sahuquillo and A. Pont, “Referrer Graph: A cost-effective algorithm and pruning method for predicting web accesses”, Computer Communications, vol. 36, no. 8, **(2013)**, pp. 881-894.

[4] M. Nykl, J. Karel, F. Dalibor and D. Martin, “PageRank variants in the evaluation of citation networks”, Journal of Informetrics, vol. 8, no. 3, **(2014)**, pp. 683-692.

[5] T. H. Haveliwala, “Topic-sensitive PageRank”, Proceedings of the 11th international conference on World Wide Web, Honolulu, United states, **(2002)** May 7-11, pp. 517-526.

[6] Z. Li, W. Yang and Z. Xie, “The summary of PageRank algorithm”, Computer Science, vol. 10, no. 38, **(2011)**, pp. 185-188.

[7] D. Zhaoyun, Z. Bin, J. Yan and Z. Lumin, “Topical Influence Analysis Based on the Multi-Relational Network in Microblogs”, Journal of Computer Research and Development, vol. 10, **(2013)**, pp. 2155-2175.

[8] W. Jianshu, L. Ee-Peng, J. Jing and H. Qi, “TwitterRank: Finding Topic-sensitive Influential Twitterers”, In Proceedings of the third ACM international conference on Web search and data mining, New York, United States, **(2010)** February 3-6, pp. 261-270.

## Authors

**LinTao Lv**, He is currently a professor of computer science and technology at Xijing University, Xi’an, China. His main research interests include complex network community discovery, information security,big data and data mining.

**QinQin Yuan**, She is currently an associate professor of computer science and technology at Xijing University, Xi’an, China. Her research interests include complex network community discovery, big data, artificial Intelligence ,data mining.