

Information tracing model based on PageRank

LI Qian^{1,2}, LAI Jia-wei², XIAO Yun-peng², WU Bin¹

(1. Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Chongqing Engineering Laboratory of Network and Information Security,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: In social network, original publisher and important nodes in the diffusion process can be traced by analyzing the spreading network of a hot topic. The participated users and spreading network structure of a hot topic build an information tracing model, which mines the source and important diffusion nodes. Firstly, it analyzed the development trend of a hot topic and extracts the users involved. Secondly, it established a user network according to the following relationship of the users involved. Thirdly, the contribution rate of users on the development of the hot topic was initialized, and the PageRank algorithm was used to construct the information tracing model. Finally, the Top k users were selected as the information publisher and important users of the hot topic according to the contribution rate. Experimental results showed that our model can effectively discover the hot topic of the publisher and important users.

Key words: social network, hot topic, information tracing, PageRank

doi: 10.11959/j.issn.2096-109x.2017.00190

1 Introduction

With the development and popularization of the Internet, human is entering a stage of information explosion gradually. In information society, the online social network has built a huge social function platform. It is using its power that social function to change people's way of life and form of society, and become a research platform for scholars to excavate the law of user behavior and the way of the information diffusion.

In order to find out the law of situational development of network group events, also to achieve the purpose of studying information traceability of network group event, this section first introduces the method of information traceability that based on PageRank algorithm in network group events. It can

be seen from Fig 1, network group event and network structure have certain relationship. A user tends to be interested in a network group and make a certain reflection, which reflects a large probability that the person is concerned about the friend, such as the user's comment or forward behavior. Whether or not to participate in the spread of the situation of the network group events, it is also related to the user influence and user interest factors. Of course, there are also random probabilities, such as watching the online group event on a hot topic and to get involved in the hot topic.

Therefore, the development and diffusion of the situation of the network group events, greatly depend on the network structure in the social network. The information source publishes the message, which is forwarded and commented by its fans, and

Received Data: 2017-06-20, **Revised Data:** 2017-07-05. **Corresponding Author:** LI Qian, liqian@cqupt.edu.cn

Foundation Items: National Key Basic Research Program (973 program) of China (No.2013CB329606), Chongqing Science and Technology Commission Project(No.cstc2017jcyjAX0099), Science and Technology Research Program of the Chongqing Municipal Education Committee(No.KJ1500425), Ministry of Education of China and China Mobile Research Fund (No.MCM20130351)

spread out in this way. This mode of information dissemination is similar to water waves.

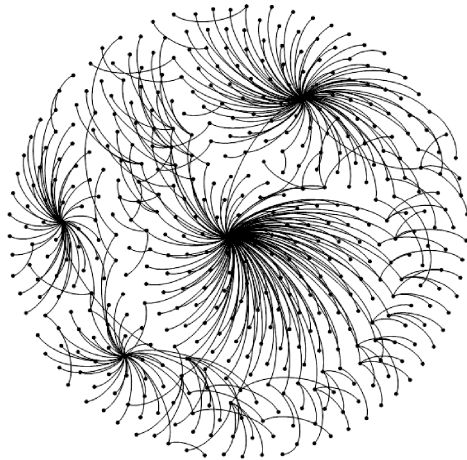


Fig.1 The classical pattern of information dissemination

The structural features of this mode of transmission can be visually seen in Fig 1: each node can be traced back to the information source according to the network structure. And the shortest path to the information source is intuitively short. According to these characteristics, many scholars have proposed the method and model of information traceability, which based on centrality. The method is that to calculate the mean and variance of all the shortest paths from other nodes to one node in a network structure, and the nodes with smaller mean and variance are estimated as the source of information. The method of information traceability based on centrality has the advantages of simple principle and easy implementation. But at the same time, the complexity of time and space is relatively high, and the calculated results tend to be locally optimal or hovering around the optimum.

The characteristics of information dissemination can be found in Fig 1, sources of information is a large contribution on the trend of development of network group events, this contribution rate can be reflected by its forwarders, the contribution rate of the forwarder can also be provided by the contribution rate of its forwarder. The method of superposition of contribution rates is particularly similar to

the thought of the PageRank algorithm. Therefore, the author designs a model of information traceability that based on PageRank algorithm in the process of analyzing and studying the information traceability of network group events.

The current research mainly study information tracing by the user role, related research for the late exploration laid the foundation. However, the pertinence and dynamic of information tracing presents new challenges for this research. This paper presents an information tracing model based on PageRank for hot topic, and establishes different information tracing models for different hot topics.

The highlights of this article are as follows.

1) Consider that each hot topic has different development trend, we build different information tracing model for every topic. According to the development trend of a hot topic, we extract users that promote the diffusion process. By leveraging the involved users, a following network is established, and contribution rate of the topic for every user are initialized.

2) Based on the PageRank algorithm, the information source model of hot topic is constructed, which is determined by the sorted contribution rate. This study can not only determine the source of the hot topic, but also discover the important users in the diffusion process.

The rest of the paper is organized as follows. Section 2 describe the related work. Section 3 formulates the problem. Section 4 explains the proposed model and corresponding learning algorithm. Section 5 presents the experimental data and analyzes the experimental results. Finally, section 6 summarizes the related work done in this paper.

2 Related work

At present, information tracing mainly focuses on evaluation of the node attributes and method based on statistical reasoning. The evaluation of node attributes is mainly to measure the network

nodes. Previous research on this aspect focused on the network structure, for example, the size of influence can be measured by degree centrality, closeness centrality, betweenness centrality and so on^[1~3]. Subsequently, some scholars have discovered the method based on the random walk, such as the Hypertext-Induced Topic Search algorithm^[4~6] and the PageRank algorithm introduced by Google^[7,8]. With the deepening of the research, the scholars found that the influence of the user is related to their behavior. There are four types of relational networks defined in Ref.[9], and mine the influence individual of the topic combined with the random walk. Ref.[10] measures the influence of the different hot topics of every user according to the concern network and the similarity of the users' interest. Tang, et al. have proposed a model of topic-factor diagram based on the network structure, the similarity of topics, users and distribution of topic information. This model can measure the intensity of the effect, and it not only can be applied to the homogeneous network, but also to the heterogeneous network^[11].

Based on the method of Statistical inference, the main reason is that the current situation of network communication comes from the infectious disease model^[12]. On this assumption, different mathematical models are used to trace the information publishers. Antulov-Fantulin et al. Presumed that the propagation status is determined by the SIR model, using the statistical reasoning framework of the maximum likelihood estimation method to trace information according to a hot topic communication network. On this basis, the researchers also proposed the AvgTopK likelihood estimation function algorithm, AUCDF estimation function algorithm and Naive Bayesian likelihood estimation function^[13, 14].

3 Problem definition

The problem to be solved in this paper as follows: confirm the publisher and the important user

nodes of the hot topics through the propagation process of the hot topic. First, a few basic concepts and related definitions are presented.

In this paper, we define the set of user relationship as follows: If user B, C, and G are the followers of user A, it is represented as B, C, and G pointing A in graph theory. Let $I(v_A)$ indicating the follower set of user A, and $|I(v_A)|=3$. In the same way, the user E and F are followee of user A, it is represented that node A points to node E and F in the graph theory. Let $O(v_A)$ represents the followee of user A, and $|O(v_A)|=2$.

To formally describe the scientific problems in this paper, we assume that V is the participated users and $E \subset V \times V$ is the relationship network of a hotspot. Let $G=(V, E)$ represents candidates and relationship network. Based on the above definitions, the inputs are candidates and relationship network. Given above input, our goal is to solve the following problems.

- 1) How to represent the contribution rate of every user?
- 2) How to determine the publisher and important users of the hot topic?

4 Model

4.1 Framework

To solve the problems above, this paper puts forward the information tracing model for the participated users of hot topic and the relationship network. The model quantifies the contribution rate of participated users for the topic development to determine the publisher and important user. The PageRank-based information tracing model framework is shown in Fig 2.

Firstly, the users involved in the hot topic are extracted through analyzing the development trend of the hot topic; Secondly, according to the user relationship, the relationship network is built. In the relationship network the user's followee as in degree and the user's following as out degree. Then, the

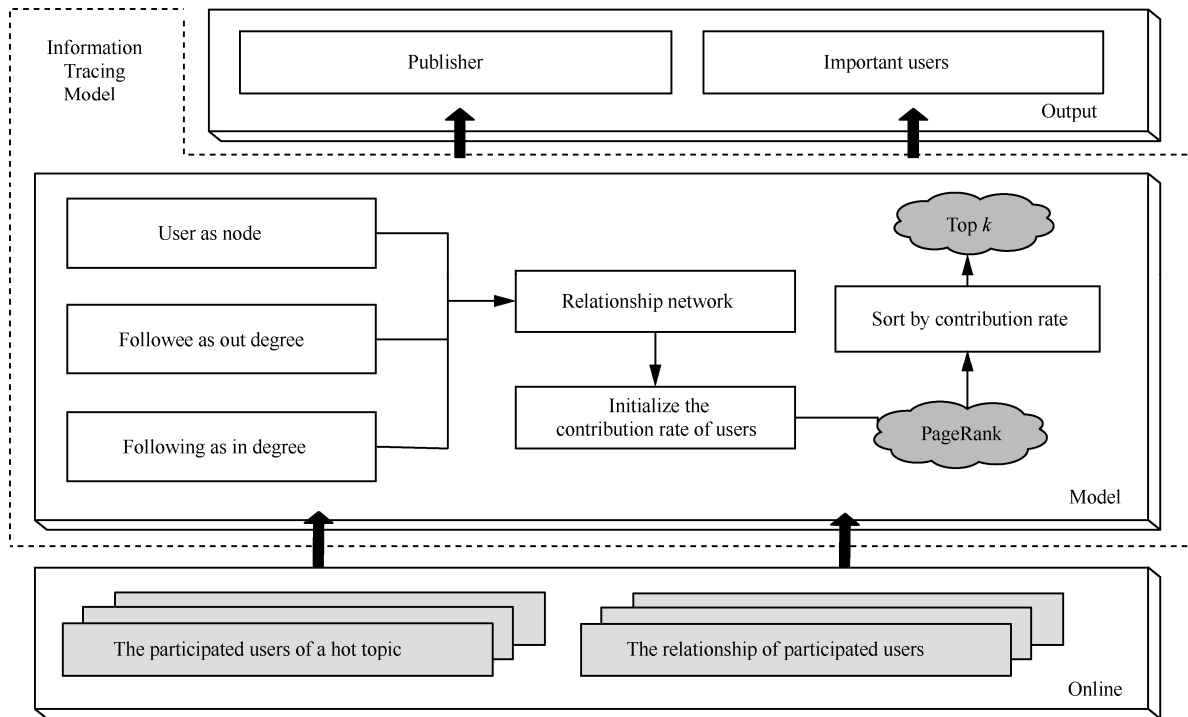


Fig.2 Model framework

contribution rate of users for the topic development is initialized. By using the PageRank algorithm, the information tracing model is established. Finally, according to the rank of contribution rate computed by the model, the publisher and important e users can be extracted.

4.2 Model detail

The model can be divided into three parts, including online acquisition, information tracing model and output, which can be seen from the overall framework of information tracing model. The development of hot topics has experienced four stages: the proposed period, the rapid development period, the peak period and the declining period. According to the topic life cycle, the basic information and relationship of users can be obtained from the social network.

The analysis for user information and relationship network shows that the development of most hot topics has some obviously common features. When a tweet is posted by a user, the following users are able to comment and forward it. Then the cascade following users also can comment and for-

ward the tweet. Then, the topic can spread like a wave.

For the first question to express the contribution of each user for the hot topic, this paper adopts the PageRank algorithm to quantify it. In the research of page rank, the PageRank value indicates the authority of the page. However, in the research of information tracing, the PageRank value means the contribution rate of users to the topic development. Further, how to build the network when applying the PageRank algorithm to information tracing? To quantify contribute of users for the development of topic, the side structure of user needs to be analyzed. In social network, user A can get the information of user B if user B is followed by user A. The direction of information flow is the opposite direction of the contribution rate flow, which is shown in Fig 3.

The message is delivered from user A to users B, C and D if user A is followed by user B, C and D. Therefore, the contribution rate of user A is represented by users B, C, and D. User A and F is followed by user G, so the contribution rate of user

G is represented by users A and F. If the contribution rate of user A is large, the contribution rate of user G will be larger. Therefore, the in degree and out degree is defined according PageRank algorithm. The followee of users is defined as out degree and the follower is defined as in degree. As shown in the Fig. 3, user A is following user G and user C, so $O(V_A) = \{v_C, v_G\}$. User A is followed by user B, D and E, so $I(V_A) = \{v_B, v_D, v_E\}$.

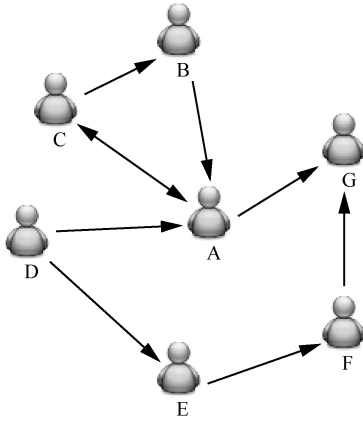


Fig.3 Relationship of users

The contribution rate is updated after initializing the contribution rate of each user. According to the PageRank algorithm, the update formula is as follows.

$$pr^{k+1}(v_i) = \sum_{v_j \in I(v_i)} \frac{pr^k(v_j)}{|O(v_j)|} \quad (1)$$

where $I(v_i)$ is the follower set of user v_i , and $O(v_j)$ is the set of follower of user v_j . $pr^{k+1}(v_i)$ is the PageRank value of user v_i , after the $(k+1)$ th iteration. Further, some users can not only get information from their followee, but also can get messages from other users in social network. Therefore, the contribution rate can also be added to a probability value the formula is as follows.

$$pr^{k+1}(v_i) = \sum_{v_j \in I(v_i)} \frac{pr^k(v_j)}{|O(v_j)|} + \frac{1}{n} \quad (2)$$

where n is the user node in the network. The first part of the above formula shows the contribu-

tion rate through the follower, and the latter part represents the contribution rate represent by any node in the network structure. The total contribution rate of user is formed of the two parts. In order to measure the importance of the two parts, the tunable parameters are introduced to control their influence according PageRank algorithm. The formula is as follows.

$$pr^{k+1}(v_i) = \alpha \sum_{v_j \in I(v_i)} \frac{pr^k(v_j)}{|O(v_j)|} + (1-\alpha) \frac{1}{n} \quad (3)$$

In order to solve second problem namely how to determine information sources user and important node users of the hot topics? Through the PageRank calculating the final PageRank value of each node, this model selects the appropriate k value to take Top k users in all user nodes. Top1 user is chosen as the information publisher of hot topic, at the same time, Top k users are important node users.

In the analysis of social network, PageRank was applied to discovery opinion leaders and evaluate influence etc. Opinion leaders discovery was oriented to all users and structure of social network, and we assume that the network is static. But in reality, the users' relationship in social network is changing all the time. In the research of a hot topic, it is necessary to study the transmission way and characteristics of this hot topic. Therefore, considering the object of information tracing model based on PageRank is a hot topic, so the established network user nodes are also the user nodes participating in the event. The relationship among them relatively changes a little for the contribution of the situation development of hot topics in the short term.

4.3 Model learning algorithm

The input of the information tracing model is the users in hot topic and its relationship network. The purpose of this algorithm is to find the publisher of the hot topic and the important user nodes in the development of the event. The information tracing model algorithm based on PageRank is described as Fig.4.

Algorithm
 information
 tracing
 algorithm
 input: Network $G(V, E)$
 Parameter α
 output: PageRank value of every user
 The first publisher and the important users.

step1: //Initialization
 step2: Initialize α .
 step3: Initialize PageRank value of every user.
 step4: **repeat**
 for each node $v_i \in V, i = 1, 2, \dots, n$
 Update PageRank value of from (3). iv
 end
until convergence.
 step5: Get sequential arrangement user base on PageRank value.
 step6: Get Top1 user to be the first publisher.
 step7: Get Topk user to be the important users.

Fig.4 Information tracing algorithm

Information tracing model is mainly divided into two parts. The first part is to calculate the contribution rate of each user in network. The second part is to confirm the information publisher and the important users through the contribution ranking. As the information tracing model algorithm described above, we can build the network with the relationship of participants. Then, initialize the parameters and the contribution rate of each user namely the

value of PageRank, and use the PageRank to update the contribution rate. Finally determine the user publisher and important transmission users. If the total number of users participating in the event is n , the operation complexity of the model is $T \sim O(N)$.

5 Experimental results

5.1 DataSet

In this paper, the experimental data collected from Tencent micro-blogging social platform. Tencent micro-blogging is one of the most popular on-line social platforms in China. In order to evaluate the information tracing model this paper takes three hot topics such as “Personal Tailor”(Topic A), “Dad where do we go”(Topic B) and ‘Panda blood girl in urgent need of help’(Topic C). We use these examples to implement the model.

Topic A is posted by “qqent”. Topic B is posted by “long story short” whose ID is “pictalks”. Topic C is posted by “liu16261”. The input of the information tracing model is the users who participate in hot topics and the relationship among them. These users are regarded as the nodes, the concerned followings and the distribution of the fans respectively as the output and the input of the chain. Fig. 5 shows the network structures build from the Topic A, Topic B and Topic C through the way mentioned above.

When the network structure constructed, the PageRank algorithm is used to realize the measurement and renewal of the contribution of each node.

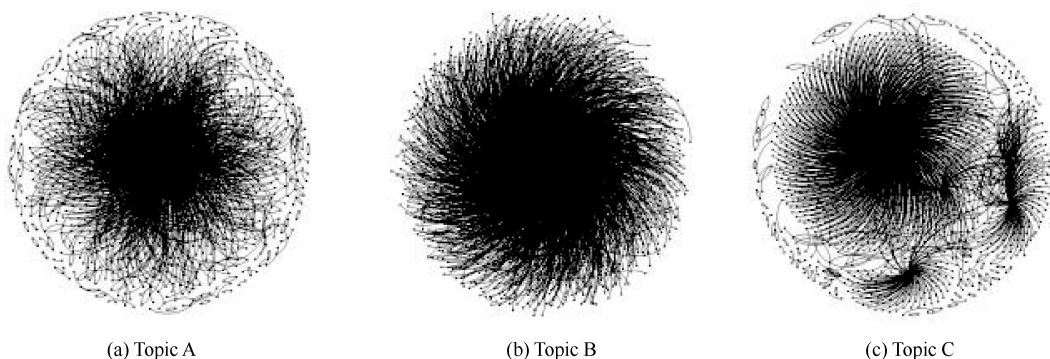


Fig.5 The disseminate structure of three hot topics

In other applications based on PageRank, it is found that the rationality of many effects is compared with the degree of entry or out of the node. For matching the results later better, Table 1 shows limited users' in-degree in three hot topics.

Table 1 Limited users' in-degree in three hot topics

Topic A		Topic B		Topic C	
Username	In-degree	Username	In-degree	Username	In-degree
qqent	377	zhangjie	1313	QQGenius	8968
cv13227	69	QQGenius	1168	sky	142
yiren7899	59	pictalks	936	weishi	126
yingzi337	38	xujinglei	652	liu16261	52
huaitu	28	nmoney	595	ph7963	20
yehuo61	24	zr19970	476	chouduo	19
zxc3855	24	QH-6577	352	weiyang	8

The purpose of the information tracing is to find the information publishers and important node users, the evaluation criteria is whether find out the information publishers and how much forward the real information publisher's rank position in all users. The metrics is mathematically defined as follows.

$$Accuracy = \left(\frac{4}{5}\right)^{L(v_0)-1} \quad (4)$$

where, $|v_0|$ is the number of all nodes in the network, is the real information publishers and $L(v_0)$ is the location meaning the real information publisher's rank sorted by the PR value through the PageRank. If the user is in the first place, the value of *Accuracy* is 1. If it is in the last place and $n \rightarrow \infty$, $Accuracy \rightarrow 0$. With the location of real information publishers more back, Accuracy reduce faster and faster, which also meet the true requirements of the information tracing.

5.2 Performance analysis

In the process of verifying the feasibility of the information tracing model with actual data, the threshold of the sum of the true value of the minimum difference is $\varepsilon = 1 \times 10^{-11}$, Fig.6 depicts the change of the value of $|PR^{k+1} - PR^k|$ in the iteration

with different α .

The Fig.6 shows that the rate of algorithm convergence is faster when α changes from 0.7 to 0.9 comparing other conditions. Based on this result, when setting $\alpha = 0.8$ through the commonly method, the rate of algorithm convergence is fastest in the Fig. 5.

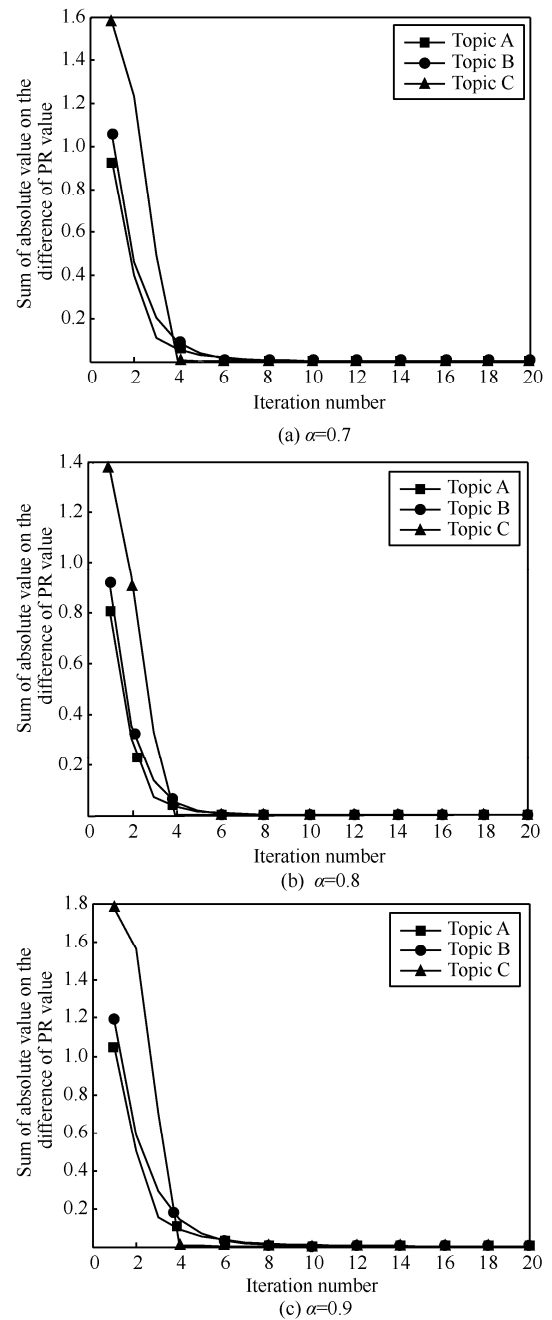


Fig.6 The value of different α

Finally, through the rank of all nodes' contribution according to the PageRank, this article consid-

ers the Top1 user as releasing the information. Table 2~4 shows the pr and number of in-links of Top10 users.

Table 2 The Top10 users' pr value and in-degree of Topic A

Username	Pr value	In-degree
qqent	0.053 75	377
cv13227	0.008 83	69
yiren7899	0.005 97	59
dongyang	0.004 15	12
huaitu	0.003 79	28
baiduxxb	0.003 72	23
yingzi337	0.003 62	38
A30783	0.003 13	17
mit49699	0.003 00	23
yehuo61	0.002 93	24

Table 3 The Top10 users' pr value and in-degree of Topic B

Username	Pr value	In-degree
pictalks	0.02119	936
QQGenius	0.01905	1168
zhangjie	0.01677	1313
xujinglei	0.00729	652
nmoney	0.00534	595
zr19970	0.00512	476
zhangyue	0.00333	241
QH-6577	0.00299	352
shouji129	0.00290	343
shoujiya	0.00233	348

Table 4 The Top10 users' pr value and in-degree of Topic C

Username	Pr value	In-degree
QQGenius	0.157 44	8968
liu16261	0.126 50	52
weishi	0.002 80	126
sky	0.001 78	142
chouduo	0.000 27	19
ph7963	0.000 21	20
hashidayin	0.000 15	7
q23002	0.000 11	7
weiyang	0.000 11	8
caozhuhua	0.000 10	5

From the Table 2~4, Topic A, Topic B, and Topic C, the number of in-links of Top10 users based on the contribution rate is large, indicating that the node users who play important role in transmitting information have been found. The publishers of Topic A and Topic B are "qqent" and "pictalks", and the actual publisher of two hot topics is also the predicted user node.

The analysis result shows that the publisher of Topic C is "QQGenius", but in fact it is "liu16261". The reason why cause such a result is "QQGenius" user has much bigger number of in-links than others when getting the number of in-links of Topic C. Therefore, it also explains a shortcoming of the information tracing model. If the gap of in-links structure is too large in information transmission, the final judgment will be affected. For the process of information transmission, if the gap of in-links is not enormous such as Topic A and Topic B, it is very suitable for information tracing model to analyze the source of information and find important node users.

Finally, the accuracy of three topics is shown in Table 5.

Table 5 Accuracy of three topics

Topic	Accuracy
Topic A	1.0
Topic B	1.0
Topic C	0.64

6 Conclusion

In this paper, a hot topic information tracing model based on PageRank is proposed, which uses the users who promote the transmission of event and the structure of social network. The model finds the source of information and important nodes of hot topics. Firstly, according to the trend of a hot topic we find these users who promote the transmission of the topic. Secondly, we initialize the contribution rate of each participating user to the development of

hot topic, and use the PageRank to construct the information tracing model. Finally, through ranking the contribution rate we find the source of information and important user nodes in Top k users. This study can be applied into recommendation systems, expert discovery, advertising and other fields. Experiments show that the information tracing model based on PageRank can predict the information publishers of hot topics and the important user nodes in the process of information transmission.

References:

- [1] OPSAHL T, AGNEESSENS F, SKVORETZ J. Node centrality in weighted networks: generalizing degree and shortest paths[J]. Social networks. 2010, 32: 245-251.
- [2] OLSEN P W, LABOUSEUR A G, HWAN J H. Efficient top-k closeness centrality search[C]//Data Engineering (ICDE). 2014: 196-207.
- [3] RIONDATO M, KORNAPOULOS E M. Fast approximation of betweenness centrality through sampling[J]. Data Mining and Knowledge Discovery, New York, 2016, 30: 438-475.
- [4] MORONE F, MAKSE H A. Influence maximization in complex networks through optimal percolation[J]. Nature, 2015, 524: 65-68.
- [5] AKOGLU L, TONG H, KOUTRA D. Graph based anomaly detection and description: a survey[J]. Data Mining and Knowledge Discovery. 2015, 29: 626-688.
- [6] BAO J, ZHENG Y, WILKIE D, et al. Recommendations in location-based social networks: a survey[J]. GeoInformatica, 2015, 19: 525-565.
- [7] MA N, LIU Y. SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork[J]. Expert Systems with Applications, 2014(3): 1357-1368.
- [8] ALEAHMAD A, KARISANI P, RAHGOZAR M, et al. OLFinder: Finding opinion leaders in online social networks[J]. Journal of Information Science, 2015, 42(9): 659-674.
- [9] ROMERO D M, GALUBA W, ASUR S, et al. Influence and Passivity in Social Media[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2011: 18-33.
- [10] WENG J S, LIM P E, JIANG J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]//The third ACM international conference on Web search and data mining. 2011: 261-270.
- [11] TANG J, SUN J, WANG C, et al. Social influence analysis in large-scale network[C]//The 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009: 807-816.
- [12] ANTULOV-FANTULIN N, LANCIC A, STEFANCIC H, et al. Statistical inference framework for source detection of contagion processes on arbitrary network structures[C]//Self-Adaptive and Self-Organizing Systems Workshops (SASOW). 2014: 78-83.
- [13] JIANG J J, WEN S, YU S, et al. Identifying propagation sources in networks: State-of-the-art and comparative studies[C]//IEEE Communications Surveys & Tutorials. 2016: 165-481
- [14] ALTARELLI F, BRAUNSTEIN A, DALL'ASTA L, et al. The patient-zero problem with noisy observations[J]. Journal of Statistical Mechanics: Theory and Experiment, 2014(10).

About the authors:



LI Qian (1980-), born in Chongqing. She is working on her Ph.D. degree at Beijing university of posts and telecommunications. Her research interests include social network analysis and information dissemination dynamics.



LAI Jiawei (1992-), born in Jiangxi. She is working on her master degree at Chongqing university of posts and telecommunications. Her research interests include social network analysis and machine learning.



XIAO Yunpeng (1979-), born in Anhui. He received his Ph.D degree of computer science from Beijing university of posts and telecommunications. He is an associate professor in Chongqing university of posts and telecommunications. His research interests include social network analysis and machine learning.



Wu Bin (1969-), born in Hunan. He received his Ph.D degree of institute of computing Chinese academy of sciences in 2002. He is a professor in Beijing university of posts and telecommunications. His research interests include complex network, data mining, massive data parallel processing, visual analysis and telecom customer relationship management.