Algorithms and Analysis of Identifying Regional Key Users of Microblog

Liang Zhang, Dan Liu, and Xu Shi

Abstract—**Today, social networking has become a popular web activity, with a large amount of information created by millions of people every day. In this paper, we focus on Sina Weibo, a rapidly growing microblogging platform, which provides a large amount, diversity and varying quality of content. Our research is from a regional perspective, and focus on the information propagation process. By measuring the real behavior of the user's blog sharing, we construct the network model of information follow and forward. We found that a small group of nodes can cover the most of the communication behavior of the information propagation. Then, we proposed a new method to rank the users, and proposed an evaluation based on real measurement coverage. Compared to similar algorithms to verify its effectiveness and accuracy.**

Index Terms—**Information propagation, microblog, social network, user identify.**

I. Introduction

The emergence of Microblogging platform Twitter, which was founded in 2006 as a symbol, according to the statistics, Twitter now has nearly 200 million users who are using 18 languages and coming from more than 120 countries in the world. China's most famous Microblogging platform namely Sina Weibo, Tencent Microblog, Sohu Microblog, Netease microblog also have got a explosive growth since then. The microblog, a new channel for the propagation of public topic, has become one of the most important self-expression, information obtaining and social practices. In the special social ecosystem of microblog, there are a relatively small group of noteworthy users, they have high reputation and influence. This part of the users plays a key role in guiding public opinion, disseminating information, forming topics, specifically the presence of a great amount of followers, and the characteristics of the broader propagation of information. They've got higher prestige and influence in the social network. To find the distribution characteristics and patterns and accurately discover these key users, is important for us to understand mechanisms underlying the formation of social topics and public feelings, the micro-effects and the social-effects made by the user behavior in the microblogging platform ,and is helpful to the guidance of the public opinion. This paper focuses on the Sina Weibo information propagation sample data capturing from a local network within a targeted university for a research purpose. Sina Weibo is currently China's most influential social networking site. It has a high popularity and access rate in the campus. Microblog users within the area of university have got many typical features, such as highly active, intensively crowded and high frequency of information exchange. We're trying to restore the actual process of the information propagation through an effective measurement and analysis upon the Microblog user information propagation network behavior data within the campus. Using this empirical and quantitative approach, we can study the propagation influence issues of the users within a targeted local region, and identify the key users of this region.

Manuscript received July 15, 2013; revised October 29, 2013.

Liang Zhang is with the National Computer network Emergency Response technical Team Coordination Center of China (CNCERT/CC), China (e-mail: zl@isc.org.cn).

Dan Liu and Xu Shi are with the International Business Machines Corporation (IBM).

II. Related Research

In recent years, a large number of domestic and foreign experts and scholars have studied on the user's propagation influence of microblog. Foreign research generally divided into three categories: First, by the measuring data. Such as Shaozhi Ye in [1]. He measured by means of the network to get the count of the user's followers, and the count of users who have forwarding or commenting the feeds to determine the user's authority level. However, this approach is relatively simple and does not consider the information propagation process and the interaction between the users, such as the following fact: the quality of the retweets or comments on the feed directly affects the quality and frequency of the propagation speed and range. Second, some scholars have proposed classic assessment algorithms such as PageRank, which was using to rank the importance of web pages. Haewoon Kwak extended the PR algorithm, he fully considered the interactions between users, and ultimately determine the user's influence [2]. Furthermore, Jiangshu Wang fully considered the different effects of different content or topic, proposed a content-based TwitterRank algorithm [3]. The study shows that common users involved in a specific topic will obtain bigger influence. The third is the spread and proliferation of user influence [4], [5]. Upon two mainstream basis, the IC and IR model simulation environment, Masahiro studied the user node's propagation influence. He used greedy algorithm to simulate the propagation process of the nodes, which represent users, finally he brilliantly coverts the user influence evaluation issue to a network node selection issue.

The main work and contributions of this paper is:

1) Microblog users use "follow" to build their social networks, this will be referred to as "following network". The propagation of information on Weibo is mainly through forwarding behavior, which will be referred to as "forwarding network". This paper will

DOI: 10.7763/IJFCC.2014.V3.279

study both the networks, compare their similarities and differences, find the forwarding network and identify the characteristics of the propagational key users in it by measuring a real network data sample.

2) Based on the core idea of the PageRank algorithm, with fully consideration the user's propagation willingness, we proposed MicroblogRegionalRank(MBRRank) as a new ranking system to evaluate the users propagation influence, and proved the effectiveness of the algorithm by the statistical data of measured samples.

III. Network Modeling and Measurement

Microblogging communication between users is based on a "following and followed" mechanism [6], which is different from the traditional social networks such as Facebook and Renren. Microblog users can follow another user without permission at any time they want, and this will make him (her) a "fan" of the user he (she) follow. Fans can obtain the be-followed user's information and any posts. This mechanism allows users to subscribe to another user's information in the most straightforward way. This fairly simple method greatly reduce the threshold for information propagation. It makes the information spreading explosively rapid.

The depth and breadth of the information propagation is not only depends on the individual's authority, the potential influence of the information itself, but also depends on the number of the followers, and the followers' followers' propagation influence. So the followers' quantity and quality, the individual's propagation influence and the willingness to spread the information is the key factor in determining efficiency of the information propagation.

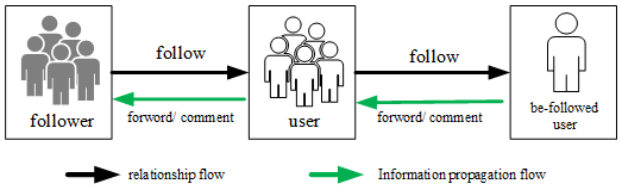

Fig. 1. Microblog users' social relationships and information propagation flow.

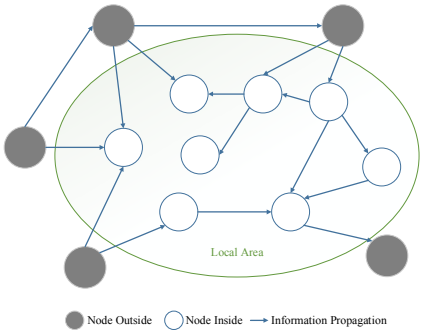

Fig. 2. Regional information propagation structure.

Information propagation structure is as shown in Fig. 2, the black nodes represent users outside the local area, while the white ones represent users inside it. In this paper we considered only the white nodes within the region. The characteristics of black nodes are temporarily without the scope.

This paper's research scope is the information propagation process, so we divided the Microblog user information propagation network into two specific network: *following network* and *forwarding network*. Traditional research paid more attention to the following network, because it is relatively represents more the natural relationships between human beings. But from our perspective, the forwarding network is the real information propagation network. It has abandoned the unnecessary interference of unwanted information, and can reflect the propagation process more accurately. So the study on the forwarding network can help us find the information propagation pattern more accurately, while the study on the following network is a valuable comparison. Related network models are defined as follows:

A. Following Network

Following Network is the interpersonal relationships formed by the user action "*follow*". Theoretically all the information published by user will be read by all his (her) followers. Following Network model describes the relationship of the user and the follower, and is the abstraction of interpersonal relationship in microblogging network.

B. Forwarding Network

Forwarding network is based upon the Following network. The user's post will only be forwarded by a small amount of followers, this constructs the information propagation process. Within an observation time window, multiple forwarding streams overlays together, this formed the forwarding network. In the forwarding network model, there is an edge between users if and if when the information forwarding relationship is presence.

IV. Data Acquisition and Analysis

The data of this research is captured from the real local network of a Chinese university. It contains all the Sina Weibo users within the local area. Students from university are highly active, intensively crowded, and have got a very high frequency of information exchange. The features of the information propagation in such area are very noteworthy. To analyze users within the area, we must firstly identify which are the region's fixed users. To solve this problem, we established a mirror machine at the core router of the campus local network, with which we captured all the campus users who had logged into the Sina Weibo platform. Our research focused on the social network structure and the interactions between users. We cared nothing about the users' private data or the content of their posts. In the data, all the information related to the specific user, is replaced by an abstract ID, all the associated timestamps are replaced by sequenced numbers. We identified a user as a campus Sina Weibo user only if at least two times of login behavior via the campus network is detected. And then the user was included in the scope of monitoring for further research. Through continuous network traffic monitoring, the information propagation route could be outlined. Since the scope of monitoring is only within a campus area, if the information is

coming from an external node, the path starts from the first campus node and ends with the last campus node, and could be called a complete propagation path. If it is original information, the propagation path start from the creator node, ends with the last campus node. If the same information comes back to the area, we consider it as a new path. In this research we captured the data from 1 May 2013 to 31 May, and observed the information propagation behaviors. The statistical data is shown in Table I:

TABLE I: THE STATISTICS OF CAPTURED DATA

Statistic items	value
Observation Time (T)	5/1/2013~5/31/2013
Microblog Post Count (N)	317789
Daily Microblog Post Count (D)	10251
Original Posts (O)	112023
Post Forwords (R)	205766
Amount of Users (W)	7990
Amount of Users Who Have Forwarding Behavior (F)	2315
Posts per User (K)	39.77

TABLE II: FOLLOWING NETWORK AND FORWORDING NETWORK

Statistic items	Following network	Forwarding network
Nodes	7990	7990
Edges	19122	3315
Isolated Nodes	2120	4076
Maximum In-degree	655	81
Maximum Out-degree	133	38
Average In-degree	8.9147	0.58
Average Out-degree	3.6398	0.29
Nodes of Maximum Connected Graph	3381	1278
Average shortest path of Maximum Connected Graph	4.353	5.971
Clustering coefficient of Maximum Connected Graph	0.3321	0.3525

The statistics show that the region has almost 8,000 active users who posted 1.28 microblog per day during the monitoring period. Users from this region have a strong willingness to use Sina Weibo for information sharing. It is noteworthy that the original posts rate is 35.25%, which shows that campus users are active at both creating and propagating information. The overall user groups in university is relatively young, they have characteristics such as active thinking, positive showing, and strong willing to share, and These characteristics may be the reason. At the same time we noticed that users who have forwarding behavior is about 28.97% of all users within the region, which indicating that the propagation of information actually benefited from a relatively small group of users. Most users who are only browsing the information, belong to the "*silent spectators*".

To take a quantitative analysis upon the regional user's microblogging behaviors, we measured and analyzed both the following network and the forwarding network. According to the previously described, there are great differences between these two fundamentally different models. Measured data for the two network model is shown in Table II:

V. KEY USER IDENTIFICATION ALGORITHM

Microblog user's propagation influence is affected by many factors. For example, the amount of his (her) followers, the amount of users he (she) followed, whether the account is certified as a "*VIP*" user, as well as the microblog itself on quality, freshness, and many other factors. Meanwhile the user's activity, frequency of logging, publishing or sharing information are also important. This paper based on PageRank algorithm, considering the willingness of the user's interaction level, proposed MBRRank algorithm to accurately assess the propagation influence of each node.

A. Algorithm Description

Given a weighted directed network $G = (V, E, W)$, nodes V, edges E, edge weight W. Edge weight W_{ij} of node i and node j represents the influence of node i to node j. In this paper, the edge weights W_{ij} is represented by *History Forwarding Ratio* $Rt(i, j)$. History Forwarding Ratio refers to the proportion that within all the be-forwarded counts that node i have how much node j shares. Defined as (1):

$$Rt(i,j) = \frac{Rtc(i,j)+1}{SC(j)+1} \tag{1}$$

$Rtc(i, j)$ is the count of node i's posts that node j have ever forwarded. Obviously, if a node forwarded node i's post ever often, it will have a much more probability to forward node i's future posts. $SC(j)$ is the count of all the forwarding made by node j during the monitoring period. Eq. (1) is the normalization of the user history forwarding statistics.

In order to measure the importance of neighbor i to j, this paper proposed the concept of *Close Ratio of Nodes*. Close Ratio is defined as the ratio of the History Forwarding Ratio of node j to node i and the sum of all the History Forwarding Ratio of nod j to the node that it have ever forwarded. In physical, it represent node i's importance for node j in all the interactions between all the neighbors of node j. Close Ratio $C(i, j)$ is defined as (2):

$$C(i,j) = \frac{Rt(i,j)}{\sum_{k(k,j)\in E^{Rt(k,j)}}} \tag{2}$$

From the definition of $C(i, j)$ we can see that, it take user's willingness to forwarding and the degree of interactions with others into consideration.

The famous Page Rank algorithm [7] is a classic algorithm that measures the degree of importance of the network nodes. Reference [1] and [3] both used this algorithm method to evaluate the microblog user's influence, and proved the effectiveness of the algorithm. The core idea of PageRank is that the PR value of each node is based on the amount of backlinks. The PR value of each node depends on all the neighbors' contributions. We apply the core idea of Page Rank algorithm to the information propagation influence assessment on microblogging network and proposed

MBRRank (MR) algorithm to evaluate the influence of each node, which defined as (3):

$$MR(i) = (1-d) + d \times \sum_{j \in B_j} MR(j) \times C(i, j) \quad (3)$$

$MR(i)$ is the MBRRank value of node i; B_i is the set of nodes that pointing to nodes i; $C(i, j)$ is the scale factor that meaning how much node j contributes its influence to node i. We use the Close Ratio defined above as this factor; d is the damping coefficient, it can be set at (0, 1). In this paper we set it to 0.87. We set the initial MBRRank value of all the nodes to 0.1, by the iteration converges, we can get all the users' MBRRank values.

B. Algorithm Comparison

To evaluate the recognition accuracy of MBRRank algorithm, we take the following three kinds of commonly used assessment methods [8] as comparisons,

1) The amount to be-forwarded (Retweets): This is the amount of the user's be-forwarded posts.
2) The number of fans (Followers): This is the amount of all the user's followers.
3) The traditional PageRank algorithm (Global PageRank): Take a directly implementation of the standard PageRank algorithm of the full sample data of the forwarding network, using the nodes degree distribution ratio as the influence spreading allocation factor.

C. Evaluation Method

In this paper, the actual influenced user coverage P is used as the evaluation indicator of user's ability of information propagation. This data is calculated by the measurement upon the count of users really influenced by the user's microblog posts within a targeted area. We summarized the count of users influenced by each post to get the final result of a user's information propagation influence. Notably, if a post is forwarded several times during propagation to user j, we counted only once. The P is defined as (4):

$$P_j = \frac{\sum_{j=[1,k]} Mc_j}{M} \quad (4)$$

Mc_j is the amount of users who were influenced by user i's j-th post; M is the total amount of nodes within the area. If a post is forward several times in the process of information propagation, all the passing nodes are counted in the scope of influence of this post.

D. Result and Analysis

In order to take further comparison and analysis of the algorithms on the evaluation indicator - the real influenced user identification accuracy, we implemented all the comparison algorithms mentioned above to get all the users' influence rank, and then compared the Top k% users' influenced-user-coverage sequence to the sequence of the actual influenced-user-coverage, analyzed the relative relationship between the sequences. The result is shown in Fig. 3:

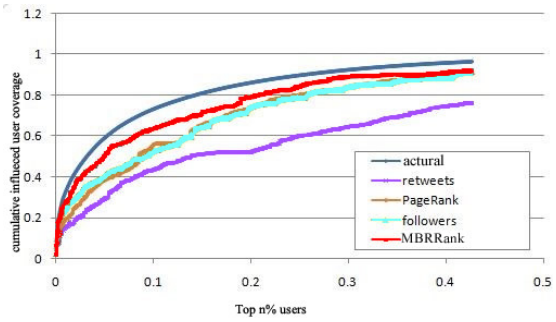

Fig. 3. Comparison for the Top n% users' cumulative coverage of the four algorithms.

Through the analysis of the data in Fig. 3, we found that Top n% users identified by MBRRank algorithm has a significantly higher coverage than other algorithms. This indicates that after the introduction of propagation willingness and degree of interactions between nodes through the Close Rate of Nodes, it can effectively approximate the user's actual propagation influence. We also found that the "PageRank" sequence and the "followers" sequence have a very similar coverage, this indicates that if the accuracy is not a noteworthy consideration, the count of followers can basically reflect the user's propagation influence. Finally we found that the "retweets" sequence has the worst effectiveness, which also confirms the research in [8]. User who has a big amount of be-forwarded posts does not necessarily means a great propagational influenced user. As can be seen from the figure, the top 20% of the users' influence is almost covering 80% of user, which meet people's everyday understanding of the 20/80 distribution.

VI. Conclusion and Future Work

In this paper we took Sina Weibo as an example, targeted in a particular area of a campus as a typical object of study, measured and analyzed the information propagation process between users, established following network model and forwarding network model. Through the comparison of measurements we found that the information propagation behavior is always driven by a small group of key users who are less than 10% of the users within the area. Then we proposed a MBRRank method to quantitatively assess each user's propagation influence. This algorithm considers both propagation willingness and degree of interactions between nodes. And finally through a measured results contrast on information propagation coverage as the evaluation indicators, we proved its effectiveness and accuracy. Subsequently, we'll study on the relationships between the individual user influence and the popular topics, consider more influencing factors in the information propagation process modeling, try to model and forecast breadth and depth of the information propagation.

Acknowledgment

We are heartily thankful to the data research team in this project, this thesis would not have been possible without their supports from the initial to the final level. This paper is sponsored by the National 242 Information Security Program

2012C98.

REFERENCES

[1] S. Z. Ye and F. L. Wu, "Measuring message propagation and social influence on Twitter.com," *SocInfo'10*, pp. 216-231, 2010.
[2] H, Kwak, C. Lee, H. Park, and S. Moon. (2010). What is Twitter? A social network or a news media? [Online]. Available: http://www. an.kaist.ac.kr/traces/www2010.html.
[3] J. Wang, E. P. Lim, J. Jiang, and Q. He. (2010). Twitter Rank: Finding topic-sensitive influential twitterers. [Online]. Available: http://www. videolectures.net/wsdm2010_weng_trft/.
[4] W. Chen, C. Wang, and Y. Wang. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. [Online]. Available: http://www. dl.acm.org/citation. cfm?id=1835934.
[5] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *Proc. the 22nd AAAI Conference on Artificial Intelligence*, 2007, pp. 1371-1376.
[6] X. G. Wang, "Empirical Analysis on Behavior Characteristics and Relation Characteristics of Micro-blog Users——Take "Sina Micro-blog" for Example," *2333e4er*, pp. 66-70, 2007.
[7] L. Page, S. Brin, and R. Motwani, "The pagerank citation ranking: Bringing order to the web," Technical report, Stanford University, 1998.
[8] H. W. Kwak and C. H. Lee, "What is twitter, Asocial Network or a News Media," *Tech Report*, pp. 591-600, 2010.

Liang Zhang was born in Liaoning in September 1980. He graduated from Peking University, major in computer system. He is a researcher working in the National Computer network Emergency Response technical Team Coordination Center of China (CNCERT/CC). He have paper in "Acta Electronica Sinica" (Vol.39, No.7, 2011,1639-1644). The title is "Test Program Generation for Microprocessor Verification Using Local Modeling Strategy", as the first author.

Dan Liu was born in Hunan in September 1983. She graduated from Peking University, major in computer architecture. She is an engineer working in the International Business Machines corporation (IBM). She has paper in "IEEE International System On Chip Conference" (2010, 182-187). The title is "TERA: A FPGA-based trace-driven emulation framework for designing on-chip communication architectures", as the first author.

Xu Shi was born in Hunan in September 1983. She graduated from Peking University, major in computer architecture. He is engineer working in the International Business Machines Corporation (IBM). He have paper in "Acta Electronica Sinica" (Vol.39, No.7, 2011,1639-1644). The title is "Test Program Generation for Microprocessor Verification Using Local Modeling Strategy", as the third author.