Empirical Research of Micro-blog Information Transmission Range by Guard nodes

Shan Chen 1,2*, Ling Ji 1, Guang Li 1,2

1School of Media Communications, LinYi University, Linyi, China
2 News and Communication Research center of LinYi University, Linyi, China

*Corresponding author e-mail: chenshan_linda@163.com

Abstract. The prediction and evaluation of information transmission in online social networks is a challenge. It is significant to solve this issue for monitoring public option and advertisement communication. First, the prediction process is described by a set language. Then with Sina Microblog system as used as the case object, the relationship between node influence and coverage rate is analyzed by using the topology structure of information nodes. A nonlinear model is built by a statistic method in a specific, bounded and controlled Microblog network. It can predict the message coverage rate by guard nodes. The experimental results show that the prediction model has higher accuracy to the source nodes which have lower influence in social network and practical application.

1. Introduction
The rapid development of mobile interconnection technology and Web technology has promoted the extensive growth of social networks in China, especially microblogs (or Weibo), which are widely used for information distribution and public opinion transmission[1]. Since 2009, the mainstream portals in China have successively introduced their own microblog products. As a major form of network services, microblogs can quickly make up inter-user online social networks with close relations and complicated structure. Thus, information can be spread quickly and widely in microblogs. However, microblogs are still new services at the development stage. Their management mechanisms are imperfect, since false or illegal information is also spread and propagated there, which will slightly hinder the economic construction and social governance in China. Thus, research on the propagation of false and illegal information in microblogs significantly underlies the monitoring, guidance and management control of network public opinions, and has critical theoretical value and practical significance. Besides, monitoring of positive information transmission in microblogs is also very significant. For instance, if a seller wants to use the business promotion and advertising media on microblogs, it is of practical significance to understand and monitor the propagation and coverage ranges of advertising information, which helps the seller to know advertising effectiveness and adjust/optimize marketing decision [2].

2. Formalized expression
2.1. Basic issues
The first issue is how to identify the range of information transmission in microblogs, which are huge social networks.
The second issue is how to determine whether one message has been transmitted to users. This issue can be determined from two basic operations (comment, repost) of this message.

For convenience, we first discuss the formalized description about analysis and evaluation of transmission coverage and degree on Weibo.

First, the information transmission in online social networks is expressed by a directed graph $G = (V, E)$, where $V$ is a set of Weibo user nodes, which are interconnected through the release, comment or repost of information; $E$ is a set of between-node edges. Then $v_i$ is the original node or transmission source point; $\overline{V}$ is a set of nodes which have received the information. $V_i$ is a subset of nodes which receive the information at time $t_i$; $v_i \in V, v_i \cap v_j = \emptyset$; the coverage rate of information transmission is defined as $O$.

**Definition 1**: The transmission coverage rate is defined as the ratio of "number of nodes who receive the information" to "number of all nodes".

$$O = \frac{|\overline{V}|}{|V|}$$

where $|\overline{V}|$ is number of nodes in $\overline{V}$; $|V|$ number of all nodes (all nodes mean the nodes in the valid user range on Weibo).

The process of information transmission is a time series $T = \{t_1, t_2, \ldots, t_i, t_{i+1}, \ldots\}$; the information coverage rate at the monitoring time $t_k$ is $O_k$:

$$O_k = \frac{|\overline{V_k}|}{|V|}$$

**Definition 2**: Guard nodes. During information transmission, if the information has been transmitted to a node subset $V_k (v_i \in \overline{V})$ at time $t_k$, or if $V_k$ is merged into the set $\overline{V}$, then the nodes in $V_k$ are called the monitoring guard nodes.

Thus, to predict information coverage rate by using Weibo guard nodes becomes to predict $O_k$ by merging $V_k (v_i \in \overline{V})$ to $\overline{V}$. Then the rules between node subset $V_k$ and coverage rate $O$ are studied, and a corresponding prediction model is built. We can estimate the coverage rate by probing the information of guard nodes belonging to $V_k$.

3. **Information coverage rate prediction model**

3.1. **Influence of transmission**

In a microblog network, each node has specific attributes and topological position, which decide its role during information transmission. The transmission influence of a node is analyzed from two levels: local and global levels.

**Definition 3**: Node influence $I$ is defined as the node degree multiplied by the average distance of indirectly connected nodes:

$$I(i) = \frac{\text{outdegree}(i) \times \sum_j d_{ij}}{\text{count}(i)}$$

where $I(i)$ is the influence of node $i$; outdegree($i$) is its outdegree; $d_{ij}$ is the distance between node $i$ and node $j$ which is indirectly connected with node $i$; count($i$) is the number of indirectly-connected nodes.

3.2. **Prediction model**

The prediction model generally works as follows: first, a model expressing the relationship between node influence and information coverage rate is built through a statistical method.

$$O(I) = f(I).$$

Based on Eq. (4), information coverage rate is estimated by analyzing whether some nodes have received this information. For instance, if the influence of node $j$ is $I_j$, then it is substituted into Eq. (4)
to estimate $O(I_j)$, which is simplified as $O_j$, meaning the coverage rate of node $j$. In practice, a group of influential nodes can be monitored in real-time. When the keywords of a message are detected, its transmission range can be estimated. For instance, the coverage rate of network public feelings and the promotion effect of Weibo advertisement can be evaluated.

Regarding the basic rules of information transmission on Weibo, when a large-influence node receives a message, the coverage rate is usually very small; but when a small-influence node receives the message, the coverage rate is usually very large. At this point, the nodes close to the source point should be excluded, because these nodes are not very influential, and when approaching these nodes, the coverage rate is usually very low. These interfering nodes can be quickly rejected so as to apply the prediction model. After detection of a small-influence node, the subsequent nodes on this transmission route will be detected. If the subsequent more influential nodes do not receive this information, this node is clarified as interfering information. Nevertheless, information coverage rate and node influence are not simply linearly related. Thus, we try a statistical method to fit a nonlinear prediction model.

3.3. Empirical analysis

As showed in the empirical disperse point distribution, the $S$-curve, which is widely applied in engineering or non-engineering computation, was selected as the basic model of regression. To more largely meet the real demands, we selected the $S$-curve in Ref. [14]:

$$y = \frac{1}{1 + e^{b + c}}. \tag{5}$$

Parameters $a$, $b$, $c$ in Eq. (5) were estimated from fitting analysis. The information coverage rate prediction model is expressed as follows:

$$y = \frac{1}{1 + 0.01e^{0.2x - 0.4}}. \tag{6}$$

The uncertain factor $R^2$ was used to determine the confidence of regression analysis. The residual errors from 20 fittings were analyzed, and the mean deterministic coefficient is 0.983. Specifically, the majority is $> 0.95$, while only a few cases of $R^2$ are $< 0.9$, indicating very high fitting confidence.

4. Experimental analysis

As showed in Table 1, in cases of a small-influence node, $\varepsilon$ indicates the effect of the prediction model, especially the results monitored from the low-influence guard nodes are very accurate. In case of a medium-influence node, the results monitored from the low-influence guard nodes are very bad, while those from the medium- and high-influence guard nodes are acceptable. In case of a high-influence node, the results monitored from the low-influence guard nodes are very good, while those from the high-influence guard nodes are not stable.

| Guard | Source point (small influence node) | Source point (medium influence node) | Source point (large influence node) |
|-------|------------------------------------|-------------------------------------|-----------------------------------|
|       | P | A | $\varepsilon$(%) | P | A | $\varepsilon$(%) | P | A | $\varepsilon$(%) |
| 3.17  | 0.987 | 0.982 | 0.51 | 0.987 | 0.921 | 7.1 | 0.987 | 0.832 | 18.6 |
| 5.67  | 0.981 | 0.975 | 0.62 | 0.981 | 0.873 | 12.3 | 0.981 | 0.761 | 28.9 |
| 10.1  | 0.952 | 0.946 | 0.62 | 0.952 | 0.914 | 4.1 | 0.952 | 0.835 | 14.0 |
| 12.3  | 0.923 | 0.914 | 1.0 | 0.923 | 0.865 | 6.7 | 0.923 | 0.827 | 11.6 |
| 19.7  | 0.755 | 0.762 | 0.9 | 0.755 | 0.752 | 0.4 | 0.755 | 0.752 | 0.4 |
| 23.2  | 0.547 | 0.556 | 1.6 | 0.547 | 0.556 | 1.6 | 0.547 | 0.556 | 1.6 |
| 24.1  | 0.526 | 0.537 | 2.0 | 0.526 | 0.531 | 0.9 | 0.526 | 0.518 | 1.5 |
| 33.1  | 0.149 | 0.147 | 1.4 | 0.149 | 0.147 | 1.3 | 0.149 | 0.145 | 2.7 |
| 36.2  | 0.08 | 0.08 | 0 | 0.080 | 0.081 | 1.2 | 0.080 | 0.090 | 11.2 |
| 40.1  | 0.08 | 0.08 | 0 | 0.081 | 0.079 | 2.5 | 0.081 | 0.080 | 1.2 |
Three experimental phenomena are analyzed here. (1) During empirical statistics, the small-influence nodes are selected as the source point and the nonlinear fitting is good, so during the experiments, the prediction model is applicable for the information transmission range of this group of source points. (2) For the small-influence guard nodes, with the increase of influence, the prediction errors are raised. The reason is that when large-influence nodes are transmitted information, the small-influence nodes will be transmitted in a fissile way. Under such situation, the selection range of the small-influence guard nodes is enlarged, leading to the decline of monitoring sensitivity. Thus, the prediction error and instability are large. (3) The prediction accuracy for the medium-influence guard nodes is always very high during the monitoring. However, since the medium-influence guard nodes are not always at the transmission path, a part of information cannot be acquired. The above experimental results are very significant for the selection of guard nodes.

5. Conclusions
We considered both global and local topological structures of nodes in Weibo networks, and used statistical methods to fit the relationship between node characteristics and information transmission coverage rate. Then appropriate monitoring nodes were selected as guards to evaluate the range of information transmission. In a controllable and bordered empirical network, data were acquired by a statistical method, and then a nonlinear fitting prediction model was built. Since this method has modest requirement and dependence on the statistical data, the prediction model is very accurate for the transmission of small-influence source points. In the future, more reasonable prediction models are needed to eliminate the effects of synchronous mechanism on the accuracy and stability of prediction.

References
[1] Yan Qiang, Wu Lianren, Zheng Lan. Research on User Behavior Characters and Mechanism in Microblog Communities[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(3): 328-333 (in Chinese).
[2] Fan Pengyi, Wang Hui, Jiang Zhihong, et al.. Measurement of Microblogging Network [J]. Journal of Computer Research and Development, 2012, 49(4): 691-699 (in Chinese).
[3] Tian JiaTang, Wang YiTong, Feng XiaoJun. A New Hybrid Algorithm for Influence Maximization in Social Networks[J]. Chinese Journal of Computers, 2011, 34(10): 1956-1966 (in Chinese).
[4] Ding Zhaoyun, Zhou Bin, Jia Yan, et al.. Topical Influence Analysis Based on the Multi-Relational Network in Microblogs[J]. Journal of Computer Research and Development, 2013, 50(10): 2155-2175 (in Chinese).
[5] Fan XingHua, Zhao Jing, Fang BinXing et al.. Influence Diffusion Probability Model and Utilizing It to Identify Network Opinion Leader [J]. Chinese Journal of Computers, 2013, 36(2): 360-367 (in Chinese).
[6] Feng Shi, Jing Shan, Yang Zhuo et al.. Detecting Topical Opinion Leaders Based on LDA Model in Chinese Microblogs[J]. Journal of Northeastern University (Natural Science), 2013, 34(4): 490-494 (in Chinese).
[7] Yu Hong, Yang Xian. Measuremet of node influence in microblogs and study of propagation path model [J]. Journal on Communications, 2012, 33(Z1): 96-102 (in Chinese).
[8] Xiong Xi, Hu Yong. Research on the dynamics of opinion spread based on social network services[J]. Acta Physica Sinica, 2012, 61(15):150509-6 (in Chinese).
[9] Xie Jing, Liu Gongshen, Su Bo et al.. Prediction of User's Retweet Behavior in Social Network[J]. Journal of Shanghai Jiaotong University, 2013, 47(4): 584-588 (in Chinese).
[10] Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter[C] // Proc of the 4th Int AAAI Conf on Weblogs and Social Medica(ICWSM’10). Menlo Park, CA:AAA, 2010:355-358.
[11] Tsur O, Rappoport A. What’s in a hashtag? Content based prediction of the spread of ideas in microblogging communities[C] // Proc of the 5th ACM Int Conf on Web Search and Data Mining(WSDM’12). New York: ACM, 2012:643-652.
[12] Yang J, Leskovec J. Modeling information diffusion in implicit networks[C] // Proc of the 10th IEEE Int Conf on Data Mining(ICDM’10). Piscataway, NJ:IEEE, 2010:599-608.
[13] Damon CEntola. The spread of behavior in an online social network experiment[J]. Science, 2010, 329:1194-1197.
[14] Zhang Wenhua, Zhang Xiang. The Suggestion for a New S-curve [J]. Mathematics in Practice and Theory, 2005, 36(6): 202-205 (in Chinese).
[15] Yang Feng, Feng Shi, Wang Lin, et al.. MICA: An Opinion Mining Prototype System for Microblog Streams[J]. Journal of Computer Research and Development, 2011, 48(Suppl.): 405-409 (in Chinese).