A Novel Method of Identifying Influential Users on Social Network

Jingchi Jiang & Wenchong Bi & Chengqi Yi
School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China

Yuanyuan Bao & Yibo Xue
Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing, 100084, China

ABSTRACT: Identifying the influential users on social network is a critical problem for advertising, promotion, public opinion analysis and network security. A few empirical studies have verified that some users could lead to wider spreading and promote the information diffusion prominently. However, the existing studies limited by a single layer analysis for information diffusion, there is still lack of a multifaceted analysis method for identifying influential nodes. In this paper, we propose a novel method for identifying the influential users in the topology of information diffusion, which contains a three-tier analytical mechanism: information-level analysis, relationship-level analysis and community-level analysis. This three-tier analytical mechanism can upgrade information-level to relationship-level, and finally identifying the influential nodes in relationship-level. The experiments on Sina Weibo demonstrate that 0.3% of users who span structural holes control 28% of the information diffusion in each event and can cause the secondary-wave of information diffusion. When the information diffuses to users with higher pagerank, more users will participate in the information diffusion.

KEYWORD: Social network; Influential nodes; Structural holes

1 INTRODUCTION

A recent report from Radicati[1] shows that social network accounts grew 14% since 2010, and reached the 3.669 million accounts in 2014. Social network such as Sina Weibo is playing an important role in China. Meanwhile, public opinion controlling, rumor spreading and advertising becoming research hotspot gradually. Consequently, there is an urgent need to analyze the effect of node, which is one of the key link in the field of information diffusion.

How to identify the influential nodes effectively and promptly in the information diffusion is crucial for advertising, promotion, public opinion analysis and network security. With the increase of information kinds on social networks, these changes make it possible to identify the influential nodes. On social networks, the process of information diffusion can be abstracted as a tree-model, in which every node is a message node. Furthermore, obtaining user relationships from every tree-model to build the users' network is the precondition of identifying the influential. If these conditions are available, we can further map the communities of these users, identifying the influential nodes from community-level.

In this paper, the main contributions are as follows:

- We propose a three-tier analytical mechanism, which includes information-level analysis and community-level analysis to identify the influential nodes in the topology of information diffusion. This mechanism can avoid the limitations of the single layer analysis.
- In order to restore the process of information diffusion, we design and implement the tree-model of information diffusion based on Sina Weibo.
- We build the user relationships from every tree-model. These users who participate in the event propagation.
- Experiments show that 0.3% of Sina users who span structural holes control 28% of the information diffusion based on the real datasets.

The rest of this paper is arranged as follows. In Section 2, we discuss the previous research on the influential nodes identification. In Section 3, we introduce the method for building tree-model and user relationships of information diffusion. In Section 4, we propose a novel method for identifying the influential nodes based on a three-tier analytical mechanism. And we also elaborate the principles of each layer. In Section 5, we conduct experiments to evaluate the proposed method. In Section 6, we conclude this paper and discuss the further work.

2 RELATED WORK

SIS (susceptible-infective-susceptible) model[2] proposed by Basu S, is the most classical model for infected disease spreading. It is used to estimate the scope of certain infectious diseases spread, it needs to give a propagation velocity, depending on the ability of the spread of disease. the spread of the propagation velocity can be represented as a node degree and the number of nodes spread function, which model can also be used in information dissemination.

Kempe D[3] presented a linear threshold model. The basic idea of this model is that if a node in an inactive state, the transition probabilities will increase as the number of active nodes increases around it. When the quantity of the node's neighbor exceeds a certain threshold value, the node converts to an active state.

David K[4] proposed model of information diffusion based on waterfall descending (decreasing cascading model), that is, when a message is propagated to the first node, which is converted to an active state - infection, and an active state node will be infected with a certain probability to activate it in a non-active state neighbors.

The influential maximization problem has been proposed and studied by Domingos and Richardson[5,8],who gave heuristics for the problem in a very general descriptive model of influence propagation.

Kempe, Kleinberg and Tardos[9] experimentally showed on large collaboration networks that the influence maximization problem in the IC and LT models, the greedy hill-climbing algorithm significantly outperforms the high-degree and centrality heuristics that are commonly used in the sociology literature.

Traditional methods for identifying the influential nodes play an important role in information diffusion. However, the traditional methods are only based on a single layer analysis, such as information-level analysis or relationship-level analysis, which will lead to the research visual angle is unitary. Our method is different from the above methods. We propose a three-tier analytical mechanism which combines information-level analysis, relationship-level analysis and community-level analysis to identify the influential nodes. This method can greatly improve accuracy and efficiency of the influential users identification, having theoretical and practical significance.

3 THE METHOD FOR IDENTIFYING INFLUENTIAL NODES IN THE TOPOLOGY OF INFORMATION DIFFUSION ON SOCIAL NETWORKS

3.1 *Three-tier analytical mechanism*

In this section, we propose a three-tier analytical mechanism to identify the influential nodes in the topology of information diffusion. The mechanism is shown as Fig 1.

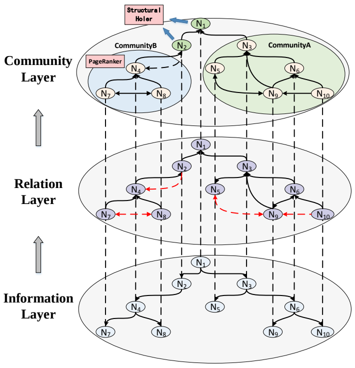

Figure 1. Three-tier analytical mechanism

Information-level. Information-level aims to restore the process of information diffusion.

Relationship-level. Relationship-level aims to map the user relations of information diffusion.

Community-level. Community-level aims to mine bridge nodes, which connect different communities and play the role of information diffusion between communities.

3.2 *Building tree-model of information diffusion*

In order to identify the influential nodes in the topology of information diffusion, we first need to restore the process of information diffusion. We abstract a process of information diffusion as a tree-model, in which every node represents a message and every edge represents a reposting chain.

We design and implement a web-parser crawler to demonstrate the feasibility of the above tree-model. The crawler can traverse the whole information reposting nodes by adopting the breadth-first strategy. Firstly, all nodes would mount the root of the tree-model. Due to the uniqueness of the message node in the topology of information diffusion, if the node in root is the same as the node in other layers we would delete the reduplicative node which is in the root and move the latter to corresponding layer.

To evaluate the above crawler, we build a tree-model of information diffusion on Sina Weibo. Fig.2 shows a real tree-model of information diffusion which *statusid* is 3736695134107027. As is shown in Fig.2, nodes represent messages and edges represent reposting chains.

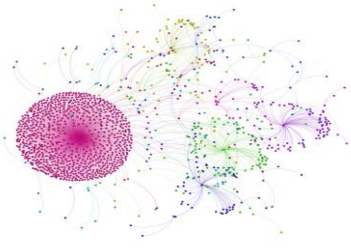

Figure 2. A real tree-model of information diffusion (statusid= 3736695134107027)

However, the tree-model can only reflects the spread of information based on information-layer and could hardly identify the influential users of information diffusion. Therefore, based on the message nodes, it is very necessary to map the users who repost messages to their relations after removal of duplicate users.

3.3 *Mapping user relations of information diffusion*

Because of the limits of API for data acquisition, the traditional method to map the user relations is pairwise comparison. However, the traditional method has a high time complexity of $O(n!)$, which n represents the number of users. Based on the above analysis, we propose a "common concerns" method to control the time complexity.

On social networks such as Sina Weibo, when user browse the web of other users, there is a special recommendation which shows the common concerns. By utilizing this feature, we build a user who follows all other users in the topology of information diffusion and map their relations by using "common concerns" method. The specific method is following:

$$\begin{cases} L_1 \rightarrow [u_1, \ldots, u_m] \\ \vdots \\ L_\beta \rightarrow [u_{(\beta-1)\bullet m+1}, \ldots, u_n] \end{cases} \tag{1}$$

$$R(u_i) = (C(L_1, u_j) \cup \ldots \cup C(L_\beta, u_j)) \cap \{u_1, \ldots, u_{j-1}, u_{j+1}, \ldots u_n\} \tag{2}$$

Eq. 1 represents that login account L executive the following tasks for user u who is participate in the event diffusion. $C(L_1, u_j)$ represents the common following between L_i and u_j. The final result $R(u_i)$ is the set of relationship between u_i and others. According to the theoretical analysis, we reduce the time complexity to $O(n)$.

To evaluate the above method, we map the user relations of information diffusion on Sina Weibo. The result of user relations (statusid= 3735192487123323) is shown in Fig.3. In Fig.3, nodes represent users and edges represent the relations.

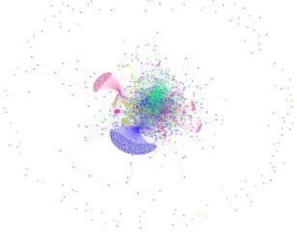

Figure 3. User relationship (statusid= 3735192487123323)

3.4 *Detecting community of information diffusion*

The purpose of detecting community is to mine bridge nodes, which connect different communities and play the role of information diffusion between communities. In generally, the exchange of information between communities is most likely to occur the second wave. Therefore, the community-level is particularly important for identifying the influential nodes.

In this paper, we use the $(\alpha - \beta)$ community detection algorithm[10] to find overlapping communities. Due to the characteristic of the algorithm, we can identify some nodes, which locate in the edge of community. These nodes are what we call bridge nodes. Though the bridge nodes don’t have higher influence, but they can make information spread more widely through the information diffusion between communities.

4 EXPERIMENTS AND EVALUTION

In order to evaluate the effectiveness and efficiency of the mapping mechanism, we conduct experiments on actual events in Sina Weibo. We choose some node characteristics including Out-Degree, Out-Degree, Average Degree, PageRank[11] and Structural Hole[12] as indicators to observe the influential users in information diffusion.

4.1 *Identifying Influential Nodes*

In order to verify the influence of different types of node in the topology of information diffusion, we select different sizes of events, including 12,821 users, 128,217 pieces of information. By adopting above three-tier analytical mechanism, we further calculate the influence of the specific nodes in information-level, and detect what kind of measurement of nodes will cause the wave of information diffusion.

The identification method is to calculate the number of child nodes, which directly connected with the specific nodes. The number of child nodes represent the number of reposted. Therefore, we can conclusion that a node has the more reposts, meanwhile, having the higher influence.

Through the above identification method, we can conclusion as shown in Fig. 4, the PageRank nodes who are detected in relation-level have the highest influence of information diffusion. Other influence ranking is in turn: Out-Degree>Average Degree>In-Degree>Structural Hole.

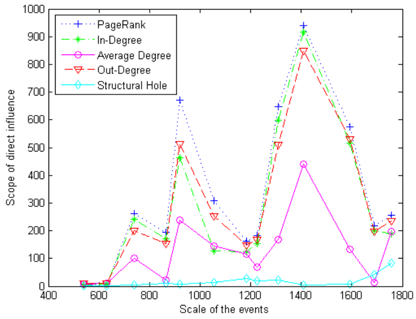

Figure 4. Comparing the influence of different nodes

4.2 *Predicting the secondary wave of event*

When an event diffuses from one community to another, it is most likely to be concerned again and we call it *the secondary wave of event*. Predicting the secondary wave of event is also important.

According to the characteristics of information diffusion, we introduce the concept of information flow. The information flow represents the process in which the information is propagated from one specific node to another. It is different from the identification of influential nodes. We calculate both the number of directly connected nodes and the number of indirectly child nodes for the specific node. We select the top 0.3% of users, which are identified by each measurement, comparing the control ability for information diffusion.

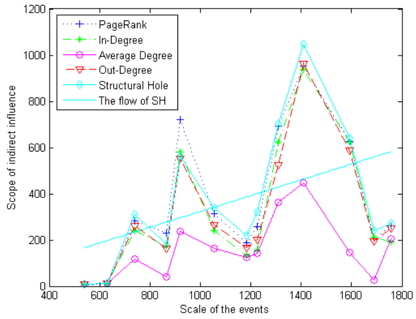

Figure 5. The comparison of the controllability

As shown in Fig 5, 0.3% of users who span structural holes have the highest number of information flow, leading to 28% of the information diffusion (reposting). Comparing the above two experimental results, we can find that structural hole spanners have less direct influence, but higher indirect influence.

5 CONCLUSIONS

In this paper, we are researching on influential nodes identifying method on social network. Based on three-tier analytical mechanism, including information-level, relationship-level, community-level, we propose a novel method for influential nodes identifying. This three-tier analytical mechanism can greatly improve the accuracy and effectiveness of influential nodes identifying. In order to testify this method, we design the Implementation method for each level. The experiments of real dataset on Sina Weibo, demonstrate the nodes with higher PageRank have the greatest influence for information diffusion, and 0.3% of users who span structural holes control 28% of the information diffusion in the average case. Our future work will focus on optimizing the algorithm of identifying and predicting the trend of events in a real-time.

ACKNOWLEDGMENTS

This work was supported by the National Key Technology R&D Program of China under Grant No.2012BAH46B04.

REFERENCES

[1] RADICATI Inc. Analyst: The second-quarter earnings of Social Network in 2014[EB/OL]. (2014-08-02). http://www.radicati.com/.

[2] Basu S, Banerjee A, Mooney R J. Active Semi-Supervision for Pairwise Constrained Clustering[C]//SDM. 2004, 4: 333-344.

[3] Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence in a social network. In: Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining. (2003) 137–146.

[4] David K, LELIS. Semi-supervised density-based clustering Proc of the 9th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2009:842-847.

[5] Domingos, P., Richardson, M.: Mining the network value of customers. In: Proc. 7th Intl. Conf. on Knowledge Discovery and Data Mining. (2011) 57–66.

[6] Goldenberg, J., Libai, B., Muller, E.: Using complex systems analysis to 2advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. Academy of Marketing Science Review(2011).

[7] Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12 (2011) 211–223.

[8] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing.In: Proc. 8th Intl. Conf. on Knowledge Discovery and Data Mining. (2012) 61–70.

[9] Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 137–146.

[10] He J, Hopcroft J, Liang H, et al. Detecting the structure of social networks using (α, β)-communities[M]//Algorithms and Models for the Web Graph. Springer Berlin Heidelberg, 2011: 26-37.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University,1999.

[12] Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 825-836.