ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 621-627
doi: 10.18178/wcse.2017.06.108

Scalable Key Node Set Mining Method in Social Network Community Based on Topological Potential and Uncertainty Measure

Hongbo Li [1], Jinbo Bai [2 +], Jianpei Zhang [3], Jing Yang [3] and Jianping Chen [1]

[1] School of Computer Science, School of Software, Zhaoqing University, Zhaoqing, China
[2] Economics & Management College, Zhaoqing University, Zhaoqing, China
[3] College of Computer Science and Technology, Harbin Engineering University, Harbin, China

Abstract. For cost savings or on-demand product promotion, e-commerce enterprises always pay great attention to key nodes in social networks. However, the exiting key nodes mining methods or indexes in networks have some deficiencies, such as not being capable of evaluating or mining communities' key nodes in networks and mining nodes playing roles of bridge between communities. Aiming at these drawbacks, based on topological potential and uncertainty measure, we propose a scalable key node set mining method in social network community. The method firstly puts the nodes in a network into two categories, the inner nodes and the boundary nodes, secondly ranks the inner nodes by their topological potential and ranks the boundary nodes by their identity uncertainty measure, thirdly searches the two ranking list respectively with parameters provided by the e-commerce enterprises, and then gets the key node set. The experiments show that the method is plausibility and validity.

Keywords: key node set, social network, topological potential, uncertainty measure.

1. Introduction

With the developing and boom of Internet, analysis of social networks, such as Twitter, Facebook, Delicious, commodity recommendation network, and so on, has become a research hot spot in academic circle [1]. Thereinto, mining communities in multifarious networks and key nodes in communities is an important research orientation. Key nodes are also named as vital nodes, influential nodes or important nodes by some researchers. Generally key nodes act as opinion leaders and authorities, and have a huge influence on other nodes [2-3].

Many researchers believe that advertising is a process to persuade audiences. Exerting influence over these opinion leader or authority nodes, we can have the effect of doing more with less. For example, if these nodes have been convinced and made a decision to buy some kind of commodity, the other nodes would most likely make a same decision, and the advertising cost would be reduced greatly. Undoubtedly this kind of method and strategy is a Gospel for e-commerce enterprises.

Though key nodes mining in social network communities is a challenging subject, we will propose a new method to mining them. The method in this paper will be based on the identity uncertainty measure, and this measure derived from the topological potential theory that can be used to effectively detective communities on all kinds of social networks.

Research of mining key nodes in networks, from small scale network to large scale network, has gone through several decades. Compared with previous works, the main contribution of this method proposed in this paper is that it can be used to mine not only key nodes with a scalable way but also bridge nodes between different communities.

+ Corresponding author. Tel.: +86 (0758)2752322; fax: +86 (0758)2752322.
E-mail address: hljbjb@126.com.

2. Related Work

Research of key nodes in networks originated in sociology, and in the process of research, scholars put forward some classic key nodes evaluation indexes, such as centrality of degree, closeness, betweenness, eigenvector, etc[4-7]. Generally the cost of the centrality nodes mining is very high, and the time complexity is $O(n^3)$ (n is the node number of the network). Though some researchers proposed some improved algorithms, the themes of the communities in the networks are ignored.

Except the classic centrality, several new indexes or mining methods of key nodes have been put forward in the recent years. Based on pioneers' works [8-9], document [10] proposed an evaluation method for node importance in communication networks. In the method the most vital nodes are defined as those whose removal with their incident links most drastically decreases the number of spanning trees. Because of this evaluation method doesn't consider interaction between adjacent nodes, so the evaluation result is not accurate. Through defining the agglomeration and combining the work of others [11-14], document [15] proposed another evaluation method for node importance. In the method the most important node is the one whose contraction result is largest increase of the network agglomeration. However, on the one hand the time complexity of the method is high, and on the other hand the method ignores the topological structure of network which leads to some special nodes, such as nodes in networks with linking structure, can't be evaluated. Document [16] proposed a method based on the neighboring nodes' node importance contribution matrix. The method initializes each node's importance value with its betweenness, and believes that the degrees of each neighboring node contribute to the importance value of the adjacent node. The main drawback of this method is its computing result doesn't accord with the reality. Different to this method, the contribution computing for adjacent node of another method put forward in document [17] includes not only the neighboring node's degrees but also their efficiency values.

With the development of IT, network scale is becoming more and more large. The common problem of the previous methods is that they are neither capable of evaluating or mining the key nodes in communities in network nor capable of mining the nodes playing roles of bridge between communities. However, these nodes often are the important marketing objects of e-commerce enterprises when they carrying on the product promotion. In addition, the previous methods can't mine key nodes with a scalable way according to the request of the enterprises when they are seriously concerning to promotion budget. Aiming at these problems, a scalable mining method for community key node set based on identity uncertainty measure will be proposed in this paper.

3. Basic Concepts

3.1. Topological potential

Document [18] provides a detecting method for network communities (hereafter referred to as the NHP method) that has a higher accuracy. Topological potential, introduced into the NHP method, is a theory derived from the "data field" in physics. In the NHP method, a network is viewed as a physical system of interacting nodes. The influence of every node is local, and the influence between the nodes is described by short-range field. Meanwhile, the NHP method stipulates that the field potential of every node is described by Gaussian potential function, and influence field of every node is topological potential field. For example, suppose network $G=(V,E)$, where $V=\{v_1,...,v_n\}$ is a nonempty finite set of nodes, $E\subseteq V\times V$ is the set of edges，$|E|=m$, then the topological potential of any node $v_i\subseteq V$ can be expressed by Gaussian potential function:

$$\varphi(v_i)=\sum_{j=1}^{n}\left(m_j\times e^{-(d_{ij}/\sigma)^2}\right) \tag{1}$$

in which d_{ij} is the distance between the nodes, measured along the shortest path. The influence factor σ is used to control the influence area of the nodes, $m_j\geq 0$ represents the mass of the node $v_j(j=1,...,n)$, and is used to described the innate attributes of every node. The influence factor σ can be determined by computing the minimum of potential entropy. In practice, σ can be optimized through stochastic search method, SA(simulated annealing algorithm), GA(Genetic Algorithm), PSO (Particle Swarm Optimization), ACO (Ant Colony Optimization). In the NHP method, potential entropy is defined as

$$H(\sigma) = -\sum_{i=1}^{n} \frac{\varphi(v_i)}{Z} \log(\frac{\varphi(v_i)}{Z}) \tag{2}$$

where $Z = \sum_{i=1}^{n} \varphi(v_i)$ is normalization factor. Based on reasonable supposition and the mathematical nature of Gaussian function, the NHP method simplifies formula (1) into

$$\varphi(v_i) = \frac{1}{n} \sum_{j=1}^{l} n_j(v_i) \times e^{-(j/\sigma)^2} \tag{3}$$

where $l = \lfloor 3\sigma / \sqrt{2} \rfloor$ is the influence area of the node v_i, $n_j(v_i)$ is the node v_i's neighboring node number at j th hop.

In the NHP method, the above-mentioned local extreme points are called representative node. For any one node $v \in V$, if there is a path leading to a certain representative node v^*, and the topological potentials of the nodes along this path increase in order, then v is considered being attracted by v^*'s topological potential. And the NHP method put the nodes in a network into two categories, the inner nodes and the boundary nodes. An inner node is the one that is solely attracted by one representative node (representative node is also considered an inner node), while a boundary node is the one that is simultaneously attracted by more than one representative node.

3.2. Uncertainty measure of the community identity of boundary nodes

In document [19] we presented an overlapping community detecting method based on topological potential that can automatically determine the number of the communities, the NSP method. Compared with the NHP method, the NSP method can not only provide a larger granularity of the communities, giving the analyst a better overview of the communities, but also introduces uncertainty measure for the community identity of the boundary nodes, and has its plausibility proved by experiments. For the NSP method, the measure is defined as

$$p_{C_i}(v) = attC_i(v) / \sum_{j=1}^{t} attC_j(v) \tag{4}$$

where $i, j = 1,...,t$, t stands for the number of communities in the network, and $attC_i(v)$ is determined by the formula below

$$attC_i(v) = \frac{1}{n} \sum_{j=1}^{l} n_j(v) e^{-(j/\sigma)^2} \tag{5}$$

where $l = \lfloor 3\sigma / \sqrt{2} \rfloor$, $n_j(v)$ is the number of the nodes that are inner neighboring nodes in the community C_i and are in the v's j th hop and the boundary nodes that can be put into community C_i. Since $attC_i(v)$ and $\sum_{j=1}^{l} attC_j(v)$ both has the factor 1/*n*, formula (5) can be simplified as

$$attC_i(v) = \sum_{j=1}^{l} n_j(v) e^{-(j/\sigma)^2} \tag{6}$$

4. Algorithm

In this paper the main steps of the algorithm are as follows.

Input: the ranking list *I* of the inner node and the ranking list *B* of the boundary node in the communities of graph *G* by applying the method in document [20], the inner key node ratio *r*, the boundary key node selecting width *w*.

Output: the key nodes set *S*.

Begin

(1)Computes the key node number k_i of the inner node in each community.

//*m* presents the community number, m_i presents the number of community node.

For *i*=1 to *m* { k_i=floor($m_i \times r$) }

(2)Computes the selecting width interval of key boundary node in each community: [0.5-*w*, 0.5+*w*].

(3)Puts top k_i inner nodes in each community to *S*.

For *i*=1 to *m* { *S*=*S*+truncate(I_i) }

(4)Puts boundary node *v* that satisfies $p_{C_i}(v) \in$[0.5-*w*, 0.5+*w*] in each community to *S*.

For *i*=1 to *m* { *S*=*S*+select(B_i) }

End

The method in document [20] deriving from the method NSP, is a community nodes importance-ranking method. It firstly divides the nodes of a community into inner nodes and boundary nodes, based on the adjacency list of communities and adjacency list of boundary nodes obtained by applying the NSP method to the network detection; secondly ranks the inner nodes by their topological potentials; thirdly ranks the boundary nodes by their uncertainty measurement values; finally combines the two ranking results. So the method NSP will get the ranking list of the inner node and the ranking list of the boundary node in the communities in a network.

In the input of the above algorithm, the inner key node ratio r and the boundary key node selecting width w reflect the requests of the e-commerce enterprises, and reflect a scalable mode, too. The taking value interval of r is (0, 1], and w is (0, 0.5].

5. Experiment and Analysis

The karate club network (as shown in Fig. 1) reflects the interaction between the members of the karate club over a three-year period [21]. Due to the tuition problems, the club split up into two smaller groups, headed by the president and the instructor respectively. Applying in the karate club network with different ratio and selecting width, the algorithm in this paper gets many key node set as shown in Table 1. In order to facilitate comparison for reader, by applying the method in document [20] the ranking list of the inner node and the ranking list of the boundary node in the communities of the karate club network are shown in Table 2, and the uncertainty measure values of the boundary nodes in the karate club network are shown in Table 3.

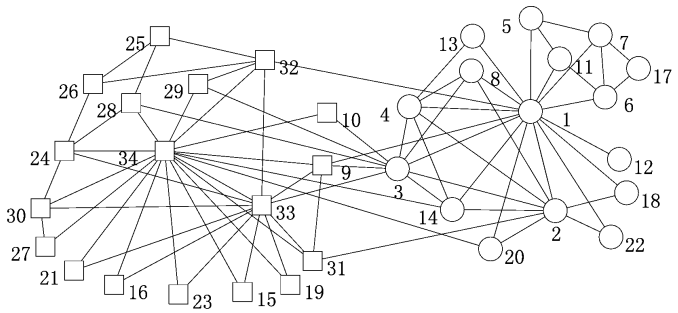

Fig. 1: Karate club network

From the Table 1 we can see that the key node set has two growing trend: with the increasing of the inner key node ratio and the boundary key node selecting width the set becomes larger and larger. These trends are completely consistent with our design purpose of the algorithm in this paper, and prove the plausibility and validity of the algorithm. Customization of the key node set size, namely scalable set, can greatly meet the request of e-commerce enterprises that spent product promotion money as required.

Table 1: Experiments data in the karate club network

No.	Ratio r	Width w	Key node set S	
			Inner key nodes	Boundary key nodes
1	0.050	0.002	{34, 1}	{}
2	0.100	0.005	{34, 1}	{3}
3	0.150	0.008	{34, 1, 33}	{3}
4	0.200	0.010	{34, 1, 33, 6}	{3, 25}
5	0.250	0.020	{34, 1, 33, 6, 24}	{3, 25, 14}
6	0.300	0.030	{34, 1, 33, 6, 24}	{3, 25, 14,20}

In the Karate club, all of the 34 members enjoy close interconnection, and their interaction is centered around the president and the instructor. Undoubtedly, the president and the instructor are the key figures in the interconnection communities. When the inner key node ratio r = 0.050, the inner key node set S = {34, 1}. The serial numbers of the nodes representing the president and the instructor are none other than 34 and 1. So the

algorithm in this paper can effectively mining the most key nodes. With the *r* value increasing, one after another node is added to *S*. From the Table 2, we can see that the new added node 33, 6 and 24 are still key nodes. Judging from this, the algorithm proposed in this paper is capable of effectively mining the key nodes from inner nodes in a network. Combining the Table 3, we can see the algorithm is capable of effectively mining the key nodes from boundary nodes, too. If a network has 2 communities, then a boundary node , which uncertainty measure value is 0.5, will possess the highest identity uncertainty. So in the experiments, with the increasing of the boundary key node selecting width *w*, the node 3 with the uncertainty measure value that is the most close to 0.5, is firstly selected.

Table 2: Ranking list of community nodes in karate club network

Community No.	Ranking list	
	Inner nodes	Boundary nodes
C_{34}	34, 33, 24, 30, 23, 21, 19, 16, 15, 27	10, 31, 28, 29, 26, 9, 32, 20, 14, 25, 3, 2, 4, 8, 18, 22, 13
C_1	1, 6, 7, 5, 11, 17, 12	13, 18, 22, 8, 4, 2, 3, 25, 14, 20, 32, 9, 26, 29, 28, 31, 10

Table 3 Uncertainty measure values of the boundary nodes in the karate club network

SN	C_{34}	C_1	SN	C_{34}	C_1	SN	C_{34}	C_1
2	0.466	0.534	13	0.305	0.697	26	0.602	0.398
3	0.504	0.496	14	0.511	0.489	28	0.664	0.336
4	0.450	0.550	18	0.329	0.671	29	0.617	0.383
8	0.425	0.575	20	0.528	0.472	31	0.674	0.326
9	0.557	0.444	22	0.329	0.671	32	0.544	0.456
10	0.679	0.321	25	0.509	0.491			

6. Conclusion

Aiming at the drawbacks of the preceding methods and indexes, such as not being capable of evaluating or mining the key nodes in communities in network and mining the nodes playing roles of bridge between communities, a scalable key node set mining method in social network community is proposed based on topological potential and uncertainty measure. The plausibility and validity of the algorithm has been proved by experiments. Therefore, this method is able to effectively mine key nodes in and between communities of networks.

7. Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61370083, 61073043, 61073041, 61402126, 71571139, 71171153), the National Research Foundation for the Doctoral Program of Higher Education of China(No.20112304110011, 20122304110012), Zhaoqing City Science and technology Innovation Guiding Category Project(No. 2015040309), Innovation and School Developing Special Fund(No. 504-20160171).

8. References

[1] P. Y. Yuan and Y. Wang. Modeling opportunistic social networks with decayed aggregation graph. *Journal of Communications*. 2015, 10 (3): 213-220.

[2] D. Jain, S. Girdhar. A study on celebrity based advertisements on consumer's purchase intentions towards selected mobile service providers in Delhi city, India. *Pranjana*. 2014, 17 (1): 9.

[3] J. D. Mart ń-Santana, A Beerli-Palacio. Magazine advertising: Factors influencing the effectiveness of celebrity advertising. *Journal of Promotion Management*. 2013, 19 (2): 139-166.

[4] U. Brandes, S. P. Borgatti, L. C. Freeman. Maintaining the duality of closeness and betweenness centrality. *Social Networks.* 2016, 44: 153-159.

[5] N. Kourtellis, T. Alahakoon, R. Simha, et al. Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*. 2013, 3 (4): 899-914.

[6] J. Putzke, H. Takeda. *Identifying key opinion leaders in evolving co-authorship networks—a descriptive study of a proxy variable for betweenness centrality*. Springer International Publishing, 2016: 311-323.

[7] L. Solá, M. Romance, R. Criado, et al. Eigenvector centrality of nodes in multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 2013, 23 (3): 033131.

[8] H. Corley, D. Sha. Most vital links and nodes in weighted network. *Operations Research Letters.* 1982, 1 (4): 157-160.

[9] E. Nardelli, G. Proietti, P. Widmayer. Finding the most vital node of a shortest path. *Theoretical Computer Science*. 2003, 296 (1): 167-177.

[10] Y. Chen, A. Q. Hu, X. Hu. Evaluation method for node importance in communication networks. *Journal of China Institute of Communications.* 2004, 25 (8): 129-134.

[11] M. O. Ball, B. L. Golden, R. V. Vohra. Finding the most vital arcs in a network. *Operations Research Letters.* 1989, 8: 73-76.

[12] L. B. Page, J. E. Perry. Reliability polynomials and link importance in networks. *IEEE Tans Reliability.* 1994, 43 (1): 51-58.

[13] L. H. Hsu, R. H. Jan, Y. C. Lee, et al. Finding the most vital edge with respect to minimum spanning tree in weighted graphs. *Info. Proc. Lett.* 1991, 39: 277-281.

[14] S. Wasserman, K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1994.

[15] Y. J. Tan, J. Wu, H. Z. Deng. Evaluation method for node importance based on node contraction in complex networks. *System Engineering-Theory & Practice*. 2006, (11): 79-83.

[16] Y. H. Zhao, Z. L. Wang, J. Zheng, et al. Finding the most vital node by node importance contribution matrix in communication networks. *Journal of Beijing University of Aeronautics and Astronautics*. 2009, 35 (9): 1076-1079.

[17] X. Zhou, F. M. Zhang, K. W. Li, et al. Finding vital node by node importance evaluation matrix in complex networks. *Acta Physica Sinica*. 2012, 61 (5): 050201.

[18] W. Y. Gan, N. He, D. Y. Li, et al. Community discovery method in networks based on topological potential. *Journal of Software.* 2009, 20 (8): 2241-2254.

[19] J. P. Zhang, H. B. Li, J. Yang, et al. Community discovery method with uncertainty measure of overlapping nodes based on topological potential. *Journal of Harbin Institute of Technology(New Series)*. 2012, 19 (2): 16-22.

[20] J. P. Zhang, H. B. Li, J. Yang, et al. An importance-sorting algorithm of network community nodes based on topological potential. *Journal of Harbin Engineering University*. 2012, 33 (6): 745-752.

[21] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*. 1977, 33 (4): 452-473.