Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Research on Influence Maximization Based on Microblog Network

Yuxin Zhou, Yufeng Liu, Huanhuan Zhi

Hunan University College of Information Science and Engineering; Changsha of Hunan 410082, China.

Abstract. Influence maximization is an important research focus in our social network. It mainly aims at searching the most influential user node quickly and accurately in the social network. Most studies are based on the quantity of the influenced parties rather than the influence propagation scope in traditional study work. A research method of influence maximization based on influence maximization community (IMMC propagation model) is proposed in the paper based on the social network relationship of microblog users. The feasibility of the method is verified from the aspects of influence quantity and influence scope through experiments of a lot of data sets.

Key words: Influence Maximization; Social Network; Overlapping Community.

INTRODUCTION

Since rapid development of WEB2.0 technology and Internet has brought great changes to our information communication in recent years. Various social networking services are produced, such as foreign large-scale social network Facebook, Twitter, and our domestic Renner, micro-blog, QQ, etc., which have infiltrated every aspect of our lives. The core function of social network is also more and more prominent with continuous increase of users and rapid increase of network scale, thereby leading to wide research, wherein influence maximization is one of key problems in the social network analysis field [1-10].

The term influence actually comes from propagation dynamic of social network in our social network. Product marketing is a direct application thereof. Influence maximization [11-13] is studied aiming at discovering finite number of influence maximization nodes as the initial active node in the social network and achieving the widest influence propagation scope. Microblog one-way attention system can abstract microblog user relationship as a directed network (the network can also be referred to as 'microblog network'). Influence maximization research also has important theoretical and practical significance for the microblog network. Our research on social network faces many challenges with continuous development of social network technology. Therefore, it is urgent to search more efficient method to solve the problem of influence maximization. Existing influence maximization research work mainly has the following limitations: 1. the potential community structure in the network is not considered, thereby affecting the propagation accuracy. 2. The important characteristics of influence propagation in the network is are ignored, including influence timeliness, acceptance rate, coverage span, etc., thereby producing certain influence on the selection of seed nodes.

A novel influence maximization research method is proposed in the paper based on microblog network (influence maximization based on influence maximization community, IMMC propagation model), main work includes the follows;

Build the microblog network according to the attention relationship among microblog users, and then explore the potential network structure from the microblog network according to the tag similarity relationship among users;

2. Implement overlapping community division on the potential network structures among microblog users and identify k influence maximization communities in the overlapping communities;

3. Measure the influence of the user node according to community influence, community module degree and attention on users, and design a formula of influence calculation;

4. Respectively calculate the influence of all nodes in the community in k influence maximization communities, and selecting the influence maximization nodes belonging to current community through comparison, thereby discovering k nodes of influence maximization from the network;

5. Design comparison experiment, prove that the proposed model can model the influence propagation process more accurately, which is better than the comparison algorithm in the scope of influence propagation.

RELATED RESEARCH

The core of influence maximization in social network [9] How to find infinite influence maximization nodes from huge network structure. Scholars in related fields have achieved very fruitful results currently aiming at research on the aspect. Richardson et al. [12, 13] firstly proposed the concept of word-of-mouth, and studied the influence maximization model based on 'viral marketing'. Kemp, Kleinberg et al. [1] firstly modeled as searching k node discrete optimization of influence maximization on propagation model. They proposed two commonly used influence propagation models: Linear Threshold model (LT) and Independent Cascade model (IC). Greedy algorithm capable of approximately reaching the optimal solution (1-1/e) is proposed. The nodes with the maximum edge benefits are selected in each round of the algorithm, which has disadvantages of high computational cost, and is not applicable to large-scale network. Leskovec [2] et al. proposed CELF (Cost-Effective Lazy-forward) algorithm in order to improve the inefficiency of the greedy algorithm, thereby reducing the seed influence scope frequency. In addition, Goyal [3] et al. proposed CELF++ algorithm, an optimized algorithm of CELF algorithm. The efficiency is improved by 35-55%, but it is also not suitable for large-scale network. Next, many efficient heuristic algorithms based on IC and LT models are proposed. For example, Chen [4] proposed MIA algorithm which is based on IC model and utilize node local tree structure to approximate influence propagation; Goral [5] et al. proposed SIMPATH algorithm under LT model. SIMPATH algorithm has good performance in running time, memory loss and influence scope. Topic factors are not considered in the above methods, thereby limiting the accuracy of seed user selection. Tang et al. [6] studied topic-wise influence intensity among users. They described that users can achieve high influence in a field generally. Caged et al. [7] studied topic perception influence maximization (TIM), calculate finite possible inquires in advance, and establish index by tree base index (INT) method, thereby effectively improving the query performance. The topic perception model proposed by Nicola et al. [8] focuses on user authority and theme interest without considering the user - user influence. The parameters of propagation model are sharply reduced, thereby improving the efficiency. The research work in the paper is different from the above research. The research mainly as the following advantages: 1. influence is studied through exploring the potential network structure among users, thereby ensuring the influence prorogation accuracy in the network. 2. The research on influence is not limited to influence node quantity, the influence scope is further considered, community structure in the network is utilized to assist us to search finite nodes with the widest propagation scope more effectively. It is more reliable to select seed nodes.

IMMC PROPAGATION MODEL BASED ON MICROBLOG NETWORK

IMMC propagation model cores mainly include discovery of k overlapping communities with the highest influence, node influence calculation based on the maximum influence overlapping community and discovery of k maximum influence nodes.

Discovery of k Overlapping Communities with the Highest Influence

The discovery process of k overlapping communities with the maximum influence is mainly introduced in the section. k Influence maximization communities are discovered through modeling microblog network and dividing overlapping communities.

Microblog Network Construction

Definition 1 (user-based attention relationship modeling): there is attention relationship among users in microblog network. Directed graph $\vec{G}(V,E)$ is used for modeling. Where in $V=\{v_0,v_1,v_2,...,v_n\}$ ($n=|V|$) refers to the set of all user nodes in the microblog network.

$E=\{e_{ij}\}$ ($r=|E|, i=0,1,2,...,n, j=0,1,2,...,n, i\neq j$) refers to attention relationship among users in the microblog network, wherein n represents the quantity of all user nodes in the microblog network, r represents the quantity of all edges in the microblog network, and e_{ij} represents that user v_j focuses on user v_i namely $v_i \rightarrow v_j$.

Definition 2 (user-based tag similarity modeling): Potential relationship among users is explored in the paper through tag information of microblog users. A potential network structure is established based on tag relationship among users, therefore we discover the potential influence maximization nodes. Undirected graph $G_{tag}(V_{tag},E_{tag})$ is used for modeling, wherein $V_{tag}=\{v_0,v_1,v_2,...,v_m\}$ ($m=|V|, m\leq n$) represents the set of all user nodes with tag information in the microblog network, $E_{tag}=\{e_{ij}\}$ ($q=|E|, i=0,1,2,...,m, j=0,1,2,...,m, i\neq j$) represents the set of tag similarity relationship between user v_i and user v_j, wherein m represents user node quantity with tag information in the microblog network, q represents tag similarity relationship quantity among user nodes, and e_{ij} represents that user v_i and user v_j have the same tag and $e_{ij} \Leftrightarrow e_{ji}$.

Overlapping Community Division Based on Microblog Network

User nodes undergo overlapping community division based on G_{tag} through DEMON algorithm [15] (Democratic Estimate of the Modular Organization of a Network, modular organization algorithm in equal identification network) in the paper.

Discovery of Influence Maximization Community

A network is constructed based on overlapping community through directed graph $\vec{G}_{overlap}(V_{overlap},E_{overlap})$, wherein $V_{overlap}=\{v_{c_1},v_{c_2},...,v_{c_s}\}$ represents set of all overlapping community nodes (each overlapping community is regarded as a node in $\vec{G}_{overlap}$), $E_{overlap}=\{e_{c_ic_j}\}(i=1,2,..s, j=1,2,...s, i\neq j)$ represents the set of all correlative relationships between node v_{c_i} and node v_{c_j}, $e_{c_ic_j}$ represents that there is a association relationship (namely community c_i and community c_j are overlapped), if node v_{c_i} and node v_{c_j} are associated, two edges $e_{c_ic_j}$ and $e_{c_jc_i}$ are respectively established, and $e_{c_ic_j} \Leftrightarrow e_{c_jc_i}$.

After overlapping community network $\vec{G}_{overlap}$ is constructed, NewGreedyIC algorithm [16] is operated for $\vec{G}_{overlap}$, thereby discovering k influence maximization community nodes.

Node Influence Calculation Based on Influence Maximization Community

After a set of k influence maximization communities is discovered through operating NewGreedyIC algorithm in the overlapping community network $\vec{G}_{overlap}$, namely $C_{\max}=\{c_i\}(k=|C_{\max}|, i=1,2,...s)$ and, it is necessary to calculate the influence of nodes included in each overlapping community within $C_{\max}$ respectively. When node

influence is calculated, factors in three aspects are considered mainly, respectively including community influence, community modularity and attention on current user node.

Definition 3 (community influence): community influence is calculated for estimating the node influence scope in the community. Current community influence is represented according to the proportion of communities affected by current community in total communities, namely:

$$F_{c_i} = \frac{num_{c_i}}{sum_{\text{overlap}}}, i = 1,2,..,k \tag{1}$$

Wherein F_{c_i} represents community c_i influence, num_{c_i} represents total community quantity affected by community c_i, sum_{overlap} represents total quantity of divided communities.

Definition 4 (community modularity): community modularity [19] is mainly used for measuring the community structure. We can know the compact degree of current communities through calculating the modularity, thereby estimating the influence propagation in the community. Modularity is higher, community polymerization degree is higher, and the influence propagation possibility in community is higher. Its definition is shown as follows:

Q_{c_i} = Community inner edge proportion - community inner edge proportion expectation

Therefore:

$$Q_{c_i} = \frac{e_{c_i}}{m} - (\frac{d_{c_i}}{2m})^2, i = 0,1,...k \tag{2}$$

Wherein $\mathrm{Q}_{c_i} \in (-\frac{1}{2},1]$ represents community c_i modularity, the value is closer to 1, the polymerization degree in the community is higher, e_{c_i} represents total edges included in the community c_i, d_{c_i} represents the sum of degrees of all nodes in the community c_i, m represents the quantity of all edges in overlapping community network $\vec{G}_{overlap}$.

Definition 5 (attention on user node): If the total out-degree of current user nodes is higher, the user has more fans, the attention on current user node is higher, and the user is more popular aiming at user node of each overlapping community in $C_{\max}$. Therefore, the attention on user nodes is defined through the proportion of all out-degree sum of current nodes in all node quantity sum in microblog network $\vec{G}(V,E)$:

$$\mathrm{A}_{v_i} = \frac{\sum d(v_i)}{sum_{\vec{G}}} (v_i \in V) \tag{3}$$

Wherein A_{v_i} represents attention on node v_i, $d(v_i)$ represents node v_i out-degree in microblog network $\vec{G}(V,E)$, $sum_{\vec{G}}$ represents the quantity of all nodes in the microblog network $\vec{G}(V,E)$.

Formula (1), formula (2) and formula (3) are integrated to calculate the community influence, community modularity and attention on user node for obtaining the influence f_{v_i} belonging to node v_i in community c_i:

$$f_{v_i} = \frac{sum_{\vec{G}}(4m^2 num_{c_i} + sum_{overlap}(2me_{c_i} - d_{c_i}^2)) + 4m^2 sum_{overlap} \sum d(v_i)}{4m^2 sum_{overlap} sum_{\vec{G}}} \quad (4)$$

Where in $c_i \in C_{\max}, v_i \in V$.

Discovery of k Maximum Influence Nodes

When we discover the set of k influence maximization communities in the overlapping community network $\vec{G}_{overlap}(V_{overlap}, E_{overlap})$, namely $C_{\max} = \{c_i\}(k = |C_{\max}|, i = 1,2,...s)$, and influence f_{v_j} is respectively calculated for all nodes $v_j \in c_i, v_j \in V$ （$j = 0,1,...,n$） in the k overlapping communities, and k influence maximization nodes can be obtained through comparing the influence of each node in each community.

Algorithm 1 is the discovery algorithm of k maximum influence nodes;

Input: set $C_{\max}$ of k influence maximization communities, overlapping community network $\vec{G}_{overlap}(V_{overlap}, E_{overlap})$, microblog network $\vec{G}(V, E)$, $P = \varnothing$.

Output: set P of k influence maximization nodes.

For all $c_i \in C_{\max}$ do,

For all $v_j \in c_i$ do,

Calculation of node v_j influence f_{v_j} ,

Nodes in community c_i are sequenced according to f_{v_j} , and influence maximization node v_j is discovered and added into set P .

End for.

End for.

EXPERIMENT AND ANALYSIS

IMMC influence propagation models proposed by the paper are respectively tested in two datasets of microblog based on IC model in the section.

Experimental Data

Two groups of datasets are used in the paper respectively from a famous star Y of Sine microblog. Data used in the paper is user relationship data obtained by the author of the paper through python Scrappy crawler frame from Sine microblog. Concrete practice is shown as follows: all users in the attention list are respectively obtained with star Y sine microblog. Then, attention list of the users and tag information are obtained, a microblog network is constructed with star Y as the starting point. The microblog network not only includes some famous stars including Y, but also contains some more popular microblog users.

Specific user names are hidden for protecting user individual information. The obtained microblog network with star S as the starting point includes 75879 nodes and 508988 edges in the network (including S node). A part of sub networks is obtained from the microblog network aiming at dataset 1 and dataset 2.

The specific statistics are shown in table 1. The same data set and relevant parameters are used for all algorithms in the comparison experiment, which do not affect the experimental results.

TABLE 1. Statistics of web datasets.

Datasets	V	E	V^*	E^*	E^*_{tag}	C	V_{com}	E_{com}
1	2699	109312	2120	86752	67769	99	At least 75	2324
2	5182	184752	4079	201783	101957	109	At least 70	5221

Table 1 shows the basic statistic data of dataset 1 and dataset 2 related networks, including node number V included in microblog network, number of edges E, number of node V^* with tags, connected edge quantity E^* composed of tag similarities between tag nodes (nodes with tags), the connected edge quantity E^*_{tag} composed of attention relationship among tag nodes, the quantity of overlapping communities C divided by networks based on tag similarities, node quantity V_{com} contained in each overlapping community, and the connected edge quantity E_{com} contained in the overlapping community network.

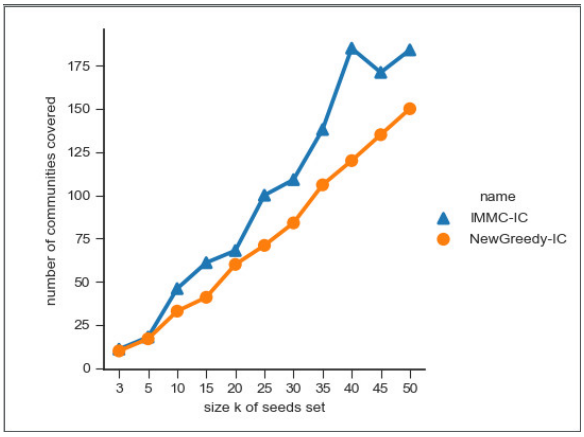
Experiment Setting and Comparison Experiment

The IMMC algorithm proposed in the paper is compared with existing NewGreedy IC algorithm. Two evaluation indexes are mainly included: one is the influence community quantity, the influence communities are more, seed user influence scope is wider, the other is influence node quantity, the influence nodes are more, and the seed user quality is better.

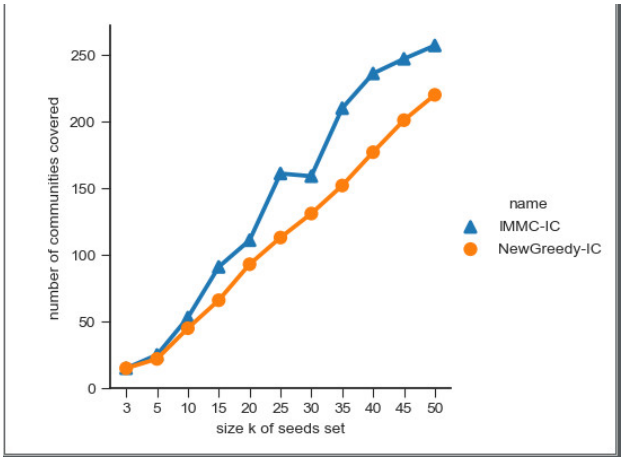
Experiment environment: Intel(R) Xeon(R) CPU E3-1226 v3 @ 3.30GHz, memory 8G; operating system: windows10; software environment: python3.5.2; tool: pharm.

Experiment and Results

The feasibility and accuracy of influence maximization research method IMMC proposed in the paper are verified through four experiments in the section. The four experiments are used for verification based on independent cascade model. Initial influence propagation probability [16] [17] p_0 is set as 0.01 in the validation process. IMMC algorithm and New Greedy algorithm are compared in the four experiments based on the above dataset 1 and dataset 2 on two performance indexes of influence community quantity and influence node quantity. Figure 1 (a) and figure 1 (b) show that the community coverage quantity of IMMC algorithm and New Greedy algorithm with seed set size of 3-50 is respectively compared in dataset 1 and dataset 2 during experiment, namely the community quantity affected by the seed set. It can be clearly observed from figure 1 (a) and figure 1 (b) that the community quantity affected by seed set discovered through IMMC algorithm is prominently higher than New Greedy algorithm with continuous increase of seed set size, thereby it indicates that the IMMC algorithm proposed in the paper guarantees the influence scope of seed node.



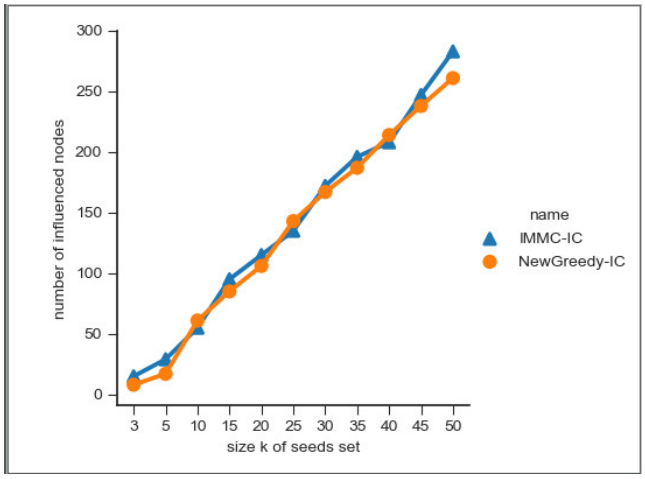
(a) Datasets 1



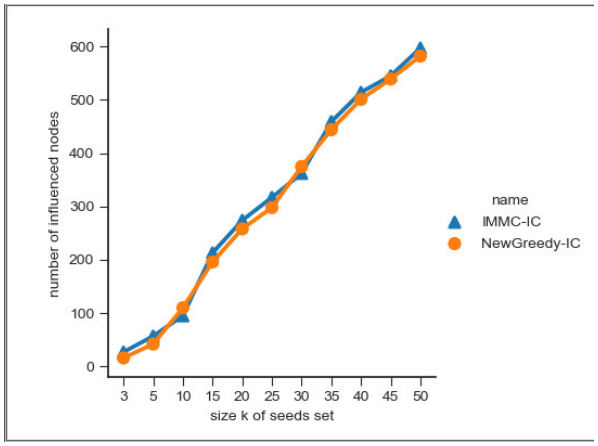
(b)Datasets 2

FIGURE 1. Number of communities covered.

The node quantity affected by IMMC algorithm and New Greedy algorithm when the seed set size is 3-50 is respectively compared in the experiment based on dataset 1 and dataset 2 in Figure 2 (a) and Figure 2 (b). The node quantity affected by the seed set generated by IMMC algorithm is more than the node quantity affected by the seed set generated by New Greedy algorithm as a whole according to Figure 2 (a) and Figure 2 (b). It is obvious that the IMMC algorithm is better than the New Greedy algorithm in the aspect of affected node quantity, thereby ensuring the influence quality of the seed node.



(a) Datasets 2



(b) Datasets 2

FIGURE 2. Number of influenced nodes.

CONCLUSION

The potential network structure among users is explored through tag similarity among users in the paper aiming at microblog user's relationship network. An influence propagation model IMMC model is proposed based on the potential network structure. An algorithm for seeking seed user set in the IMMC model is designed, thereby assisting us to position k seed users of influence maximization more accurately in the microblog network, and the widest influence scope of k discovered seed user nodes can be guaranteed. The IMMC model is experimented in the microblog dataset, which is composed with the traditional New Greedy algorithm. The seed user node obtained through the IMMC algorithm can reach excellent influence effect in the aspects of influence width and influence depth, thereby verifying the accuracy of IMMC algorithm. However, the study still has defects, namely the time complexity of the algorithm may be relatively high. Therefore, the influence calculation should be further optimized in subsequent study to reduce the influence calculation time, thereby designing more efficient initial seed set selection method.

REFERENCES

1. Kemp D, Kleinberg J, Taros É. Maximizing the spread of influence through a social network[J]// Proceeding of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C. August 24-27, 2003.USA:ACM, 2003:137--146.
2. Leskovec J, Krause A, Gastrin C, et al. Cost-effective outbreak detection in networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007:420-429.
3. Goal A, Lu W, Lakshmana L V S. CELF++: optimizing the greedy algorithm for influence maximization in social networks[C]// International Conference Companion on World Wide Web, Hyderabad, India, March 28 - April 01, 2011. USA: ACM, 2011:47-48.
4. Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010. NY, USA: ACM, 2010:1029-1038.
5. Goal A, Lu W, Lakshmana L V S. SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model[C]// IEEE: International Conference on Data Mining, Vancouver, BC, Canada, 11-14 Dec. 2011. USA: IEEE Computer Society, 2011:211-220 [6] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28-July 01, 2009.NY, USA: ACM, 2009:807-816.
6. Alay C, Beriberi N, Bronchi F, et al. Online Topic-aware Influence Maximization Queries[C]// ETBE: extending database technology, Athens, Greece, March 24-28, 2014.Proc. 2014:295-306.

7. Beriberi N, Bronchi F, Minco G. Topic-aware social influence propagation models [J]//ICDM, 12th International Conference on Data Mining, Brussels Belgium, Dec 10-13, 2012. Verilog London: Springer, 2013, 37(3):555-584.
8. Chu Y, Zhao X, Liu S, et al. An Efficient Method for Topic-Aware Influence Maximization [M]// APWeb2014: Web Technologies and Applications, Changsha, China, September 5-7, 2014.Switzerland: Springer, 2014:584-592.
9. Zhang H, Nguyen D T, Zhang H, et al. Least Cost Influence Maximization across Multiple Social Networks [J]. IEEE/ACM Transactions on Networking,2016,24(2): 929-939.
10. Brown J, Renege Social ties and word-of-mouth referral behavior[J] Journal of Consumer Research, 1987, 14(3): 350~362.
11. Richardson M, Domingo's P. Mining knowledge-sharing sites for viral marketing[C]// KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, July 23-26, 2002. New York: ACM, 2002: 61-70.
12. Cao X, Yu Y. Assents: A Benchmark Dataset of Aligned Social Networks for Cross-Platform User Modeling[C]// ACM International on Conference on Information and Knowledge Management, Indianapolis, Indiana, USA, October 24-28, 2016. NY, USA: ACM, 2016:1881-1884.
13. Goo W, Wu S, Wang L, et al. Social-Relational Topic Model for Social Networks[C]// ACM International on Conference on Information and Knowledge Management, Melbourne, Australia, October 18-23, 2015.USA: ACM, 2015:1731-1734.
14. Michele Cassia, Giulio Rossetti, Fiscal Gannett, et al. DEMON: a local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012, 615-623.
15. Wei Chen, Yanjun Wang, Siu Yang. Efficient maximization in social networks[A]. In proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France, 2009.NewYork: ACM NewYork,2009: 199-208.
16. D. kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network[C], in Proceedings of the nineth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.137-146,2003.
17. G. Nemhauser, Wolsey, and Fisher. An analysis of the approximations for maximizing sub modular set functions [J]. Mathematical Programming, 14(1):265-294, 1978.
18. Newman M E Girvan Minding and evaluating community structure in network [J]. Physical Review E, 2004, 69(2):026113.