International Journal of Hybrid Information Technology
Vol.8, No.7 (2015), pp.399-410
http://dx.doi.org/10.14257/ijhit.2015.8.7.36

A New Multi-View and Controllable Community-Uncovering Algorithm

Zhang Xin, Wang Xiaolong and Liu Bingquan

Harbin Institute of Technology, school of Computer Science and Technology, Harbin, 150001 China
xzhang@insun.hit.edu.cn

Abstract

This paper introduces a new multi-view and controllable community-uncovering algorithm, an achievement of improving PageRank algorithm and Spin-glass model, which can avoid the overlapping community structure in the process of detecting communities by means of other algorithms and also helps to improve the usual community-expansion model. The process of uncovering communities by the introduced algorithms can be divided into three steps: first, identifying the nuclear one among nodes ranked by the advanced PageRank algorithms; Second, through using multi-view recognition modularity provided by Potts spin-glass model, optimizing the expansion model of local community that is found by applying the improved Iterative Greedy algorithm to eliminate the traditional modularity's limits in the resolution limit and the following negative effects. Finally, grasping the overlapping structure and notes carefully. By analyzing and comparing the two results of respectively using PRSGMFCA and traditional technical schemes in both computer simulation network and the real network, it proves that the former enjoys stronger stability and higher accuracy than the latter, and its computation complexity is also acceptable.

Keywords: *Improved PageRank algorithms, Spin-glass model, Multi-view recognition and controllable, Iterative Greedy algorithms, Community-uncovering*

1. Introduction

Social network reflects the social developing law. Analyzing its relevant activities and their laws has both important theoretical and practical meaning in promoting our social environment's healthy and continuous development, effectively resolving urgent events and shaping a good social morality. Now there are some studies about the local communities in microblog. Reference [1,2] introduces TwitterRank algorithm[3], created by analyzing Twitter users' homophily and advancing PageRank web page's weight, which because its ability to calculate users' weight can uncover the most dynamic user group in the microblog. Mr Wu and his colleagues give us a new algorithm-XinRank algorithm developed from the advanced TwitterRank algorithm [4]. And then they use their algorithm to rank the users of Sina micrlog [5], China, according to their importance. Community reflects the characteristics of network users' behavior and the correlation between two users [6]. Researching network community is crucial to get a clear eye on network function and its structure, and also makes it easier for people to detect the relationship of network participants. In Internet social apps Web Community is very common [7]. For example, using community uncovering algorithm in microblog can improve the effects of advertising campaign; and e-commercial users can use community-uncovering algorithm to build a more stable and precise recommended system, because by using this algorithm to study the records of searching, they can do a

ISSN: 1738-9968 IJHIT

research and make a conclusion about the users' behavior, then as a result providing a more satisfactory searching results to the app users. However community-uncovering technology still has many defects such as community localization and community overlap. To solve the problems above, some scientists have gave us some algorithms such as CPM [8], GCE [9], LFM [10], MONC [11]. Although these algorithms referred before can to some extent diminish the problems in uncovering local community and overlapping community, there are still many problems left to us: ① node shaking. The peers in the network joining and leaving frequently makes algorithm unable to finish its calculation; ②repetitive calculating. The algorithm is too complex to avoid calculating fitness value repeatedly; ③ Linear expansion. Web Community grows only in a certain direction, breeding a "linear community" trend clearly.

Creative Ideas:

(a)In order to effectively eliminate the above problems, this paper provides a new way: choose the nuclear one as the seed node, selected from nodes sorted by following Pareto effect rules in the network and using the advanced PageRank algorithm, to improve the ability to uncover the local community.

(b) Using multi-view recognition modularity based on Potts spin-glass[12] to optimize the local community expansion modularity can avoid the limits in the resolution limit of the normal modularity and the following negative effects; Uncovering local community structure in guidance of the local information sent by the nuclear node enables this algorithm to solve the problem of overlapping community easily; Taking into account the distinctive characteristics of the local community, Iterative Greedy Algorithm is chosen to uncover such community , so that to detect successfully the overlapping structure and nodes finally.

2. Algorithms

A. The Target Function of the Local Community Expansion

In this paper, Hamiltonian, explained clearly in his study by Reichardt, is the target function of the local community expansion, represented by formula 1 below:

$$\begin{aligned} H(\mathrm{C}) = & -\sum_{i\neq j} a_{ij} A_{ij} \delta(\mathrm{C}_i, \mathrm{C}_j) \\ & + \sum_{i\neq j} b_{ij} (1 - A_{ij}) \delta(\mathrm{C}_i, \mathrm{C}_j) \\ & + \sum_{i\neq j} c_{ij} A_{ij} (1 - \delta(\mathrm{C}_i, \mathrm{C}_j)) \\ & - \sum_{i\neq j} d_{ij} (1 - A_{ij})(1 - \delta(\mathrm{C}_i, \mathrm{C}_j)) \end{aligned} \quad (1)$$

In the formula above, "m" refers to the total number of margins and "c" as collection of community. More is the number of margins between two spinning nodes with the same direction and speed, fewer is the number of margins between two different spinning nodes and the higher is Hamiltonian. a_{ij}, b_{ij}, c_{ij} or d_{ij} respectively represents a weight. And here, the spinning nodes with the same direction and speed are in the same community. When $a_{ij} = c_{ij} = 1 - \gamma \frac{d_i d_j}{2m}, b_{ij} = d_{ij} = \gamma \frac{d_i d_j}{2m}$, Hamiltonian can continue to be calculated in the formula 2 below:

$$H(C) = -\sum_{i\neq j} (A_{ij} - \gamma \frac{d_i d_j}{2m}) \delta(C_i, C_j) \quad (2)$$

In this paper"2H(C)"is the target function of local community-uncovering.

B. Using the Advanced PageRank Algorithm to Choose the Nuclear Node

Here, because a local community is treated as a collection of a potential leader and its followers, uncovering the nuclear node, which is regarded as the seed of local community, can make local community-uncovering easier and more accurate.

Here, the advanced PageRank algorithm is chosen as the measurement, which is used to rank all network nodes. By this way we can find the nuclear node. And in order to make PageRank algorithm applicable in the ranking of nodes in the indirect graph, we must optimize the algorithm.

Definition 1 (Centrality). "G=(V, E, w)" represents an indirect weighted network, "w" weighting function, PRcen (i) the centrality of node and it can be calculated by the formula 3 bellow:

$$PRcen(i) = c\sum_{j\neq i} PRcen(j)\frac{w_{ji}}{\sum_{k\in adj[i]} w_{ik}} + \frac{(1-c)}{N} \quad (3)$$

In the formula above, "N" represents the number of G's nodes; "c" is a constant and its range can be represented by " $c \in (0,1)$ "; $adj[i]$ refers to the collection of all approximal points of "i". Here the limit of the range of "c" can not only accelerate the convergence of the algorithm, but also is helpful for the convergence caused by isolated nodes. In general, "c" is about 0.85. And "τ "is usually changed with the specific need. What is described above is the traditional PageRank algorithm, which is optimized here, aiming to enable it to use the link information of a indirect graph network and to represent "*PRcen* "more accurately and timely. In the advanced algorithm, recursion is fitted in the specific condition here, and the centrality of every node is determined by the centrality of its approximal node. And every node's weight is calculated by the sum of weight of node and its approximate node multipling a certain proportion.

In other words, "Prcen" is used to described the centrality. The higher the centrality of a node is, the higher the centrality of its approximal node is. Weight is used to represent the strength of the junction between two nodes. The higher the weight is, the higher the relative centrality is. The formula for calculating it is as below:

Algorithm 1:

Input: τ, c

Output : $PRcen$

1.$\forall i, PRcen_s[i] \leftarrow 1/N$

2.*while*$(res > \tau)$ *do*

3.$\forall i, PRcen_t[i] \leftarrow 0$

4. *for all i do*

5. *for all j do*

6. $PRcen_t[j] \leftarrow PRcen_t[j]$
$+PRcen_s[i] \times (w_{ij} / \sum_{j\in adj[i]} w_{ij})$

7. *end for*

8. *end for*

9. $\forall i, PRcen_t[i] \leftarrow c \times PRcen_t[j] + (1-c) \times (1/N)$

10.$res \leftarrow \|PRcen_s - PRcen_t\|$

11.$PRcen_s \leftarrow PRcen_t$

12.*end while*

The first several nodes in the rank are possible to be chosen as the nuclear node. Usually initialization is to a large extent decided the algorithm for dividing the network,

so in algorithm 1 the algorithm can finish expansion rapidly if starting calculating with a correct nuclear node, but by contrast, if starting with the wrong nuclear node the result would be repeated, iterative and invalid. In order to avoid the problem algorithm 1 is an unusual community-uncovering technology, which has a beneficial effect on initialized enumeration and a high stability. The algorithm also largely reduces the possibility of emerging redundant community.

C. The Multi-view Recognition Community Expansion Model

Multi-view Recognition Modularity

Definition 2 (Multi-view Recognition Modularity): To get rid of the influence and restriction of resolution limit put forward by Newman, the author introduces the parameter γ to measure the community scale and tries to find a suitable algorithm for local community by optimizing Hamiltonian based on the Potts spin-glass Model. This algorithm can be called Multi-view Recognition Modularity, described in the following Formula 4:

$$mQ(C) = \frac{1}{m}(m_c - g\frac{k_c^2}{4m}) \quad (4)$$

In Formula 4, C represents a network community, m the total margins, m_c the inside margins of C, k_c the total nodes and γ the regulate parameter of multi-view recognition($0 \le \gamma < +\infty$) .When a node joins the community C, the modularity formula transforms into the following Formula 5:

$$mQ(C \cup v) = \frac{1}{m}[(m_c + d_{v_c}) - \gamma\frac{(k_c + k_v)^2}{4m}] \quad (5)$$

In Formula 5, d_{v_v} represents the margins linked v with nodes in C.Therefore, after new node's joining, the modularity's variable value is:

$$\begin{aligned}\Delta mQ(C,v) &= mQ(C \cup v) - mQ(C) \\ &= \frac{1}{m}[(m_c + d_{v_v}) - \gamma\frac{(k_c + k_v)^2}{4m} - m_c + \gamma\frac{k_c^2}{4m}] \quad (6) \\ &= \frac{1}{m}(d_{v_c} - \gamma\frac{k_v^2 + 2k_c k_v}{4m})\end{aligned}$$

By introducing new nodes and solving mQ, the original nuclear node can be expanded and generally it can be developed into a local community. After inputting a local community C, in the formula solving mQ will output a real value which reflects the density and linkage between the internal and external margins of community. If the different external nodes are introduced into the community, the value will increase or decrease. Here we will choose the node that contributes to the highest increase, so we use the iterative method to search the nodes that devotes to the increase of value until we find the highest one, which can fulfill the expansion of the community C .There is one point to emphasize that it is the parameter γ that protects the mQ from the influence and restriction of resolution limit. Meanwhile, every local community expands so separately that what we finally find is natural overlapping community.

Improved Hierarchical Greedy Expansion Model

In reality, the local community's expansion meets many problems, including repeated calculation and node shock that described in the introduction. To solve such problems, the optimized hierarchical greedy expansion model is introduced. The principle is using an original nuclear seed node to obtain the local community structure predicted from the following Algorithm 2.

Algorithm 2:

2.*L*ocal Expansion

Input: leader v,parameter γ

Output : *A local community C*

1.$C \leftarrow v, S \leftarrow \phi, l \leftarrow 0$

2.$S \leftarrow N_i^{(l+1)} - N_i^{(U)}$

3.$\forall v_i \in S, v_i.\Delta \mathrm{m}Q \leftarrow \Delta mQ(\mathrm{C}, \mathrm{v}_i)$

4.$\mathrm{v}_{\max} \leftarrow$ *node with* $\max(\Delta \mathrm{mQ})$

5.if $v_{\max}.\Delta mQ < 0$

7. *return C*

8.*end if*

9.*while* $S \neq \phi$ *do*

10. $v_{\max} \leftarrow$ *node with* $\max(\Delta \mathrm{mQ})$

11. if $v_{\max}.\Delta mQ > 0$

12. $C \leftarrow C \cup v_{\max}, S \leftarrow S - v_{\max}$

13. $\forall v_i \in S, v_i, \Delta mQ(\mathrm{C}, v_i)$

14. end if

15. else

16. $S \leftarrow \phi$

17. *end else*

18.*end while*

19.$l \leftarrow l+1$, *goto line* 2

3. Experiments and Performance Analysis

To measure the performance of the algorithm referred above, the test needs to be performed in the computer simulated network and real network. Here the detailed introduction of the computer used for calculation is as follows:

HP PC Intel(R) Pentinim(R) CPU P600

@1.87GHz 4G RAM Microsoft Windows 7 OS .

The program the computer used is Java 6.0. The basic tested data are from the computer information based on LFR-benchmark and some information based on the real network, while the analysis algorithms used in the test are CPM [4], GCE [5], LFM[6] and FN[12] referred above.

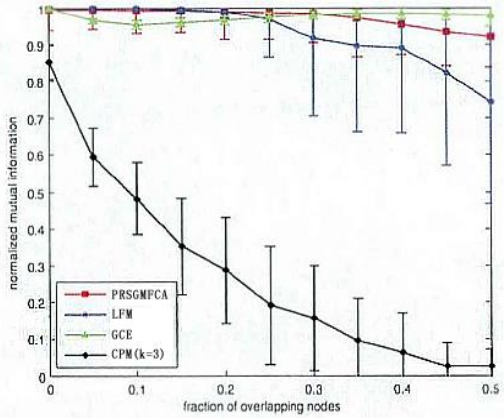
A. The Experimental Performance Analysis of the Simulated Network

We use the LFR-benchmark method to build the computer simulated network. The building process involves the distribution characteristic between the node and the community and on the other side it will devote to the hierarchy and overlap among communities. During the experimental process, we set the network number as 50 based on these parameter settings specified in Table 1 while we use the NMI (Normalized Mutual Information) to evaluate the experiment results described in Diagram 1(Diagram (a) and

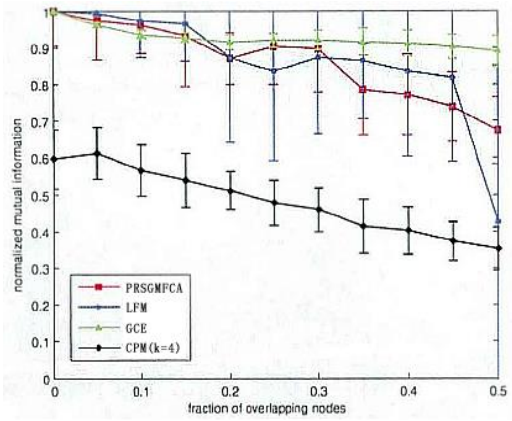
(b) represents the results of the use of PRSGMFCA, CPM and GCE in the simulated network G1 and G2, while Diagram (c) and (d) represents the running results of the PRSGMFCA, FN and CPM in small community network G3 and G4. The error line shows the average error value when the algorithm runs 50 times).After comparing (a) and (b), we find that in a network (m=0.1) with clear structure most NMI of PRSGMFCA exceed 0.9 and have an excellent stability, while the stability of LFM is not good, for it chooses seeds so arbitrary that the results are obvious different. However, the network G2 (m=0.3) looks sparse and the internal margins of it are not denser than that in G1 while the margins between communities increase obviously. It also means that when a community structure is not so clear, the effectiveness of its algorithm will weaken along with the increasing of overlapping nodes. And CPM algorithm is easy to be affected by k (referred as the size of clique here), so the result calculated by it is not perfect. However, when the on value stays at a large number, the detecting of community structure from crowded overlapping communities performs well. But the use of improved PRSGMFCA is superior to other algorithm (CPM, LFM and GCE), which means it is very necessary to improve the strategies of choosing the original nodes and relative algorithms. On the other side, from the Diagram (c) and (d) we find the FN algorithm can not be detected structure effectively when the sizes of community G3 and G4 are lower than the recognized lower limit of Newman modularity, that is $\sqrt{2m} \approx 141$.And we can also noticed that for most parameters the results calculated by PRSGMFCA are about 0.9, which means that this improved model is not restricted by the resolution limit.

Table 1. Parameter Settings of the Simulated Network

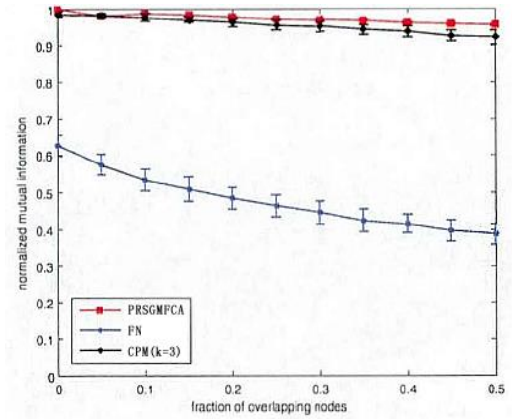
Net work	N	k	k_{max}	C_{mix}	C_{max}	t_1	t_2	m	O_n	O_m
G1	1000	20	50	20	100	−2	−1	0.1	0 −500	2
G2	1000	20	50	20	100	−2	−1	0.3	0 −500	2
G3	10000	20	50	20	100	−2	−1	0.1	0 −500	2
G4	10000	20	50	20	100	−2	−1	0.3	0 −500	2



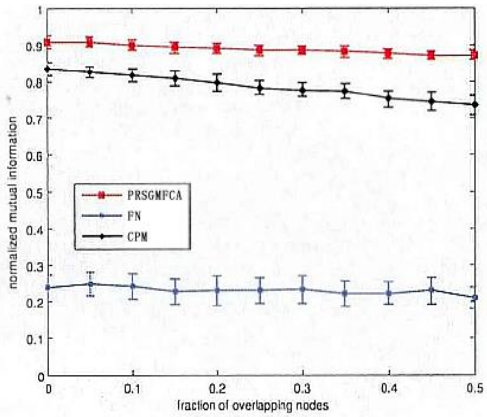
(a)Experimental Result Diagram of Network G1



(B)Experimental Result Diagram of Network G2



(C)Experimental Result Diagram of Network G3



(D)Experimental Result Diagram of Network G4

Figure 1. Experimental Result Diagram of Simulated Networks

B. The Experimental Performance Analysis of the Real Network

In addition to the condition discussed above, we also need to find some real network data to test the effectiveness of improved model. So we introduce several real experimental networks, such as Zachary's karate club, Dolphins' social network, Books about US politics Books and American College football described in Table 2. Meanwhile, the distribution of communities will be assessed by expanding modularity (EQ) described in Formula 7.

$$EQ = \frac{1}{2m}\sum_{C}\sum_{i,j\in C_k}\frac{1}{O_i O_j}[A_{ij} - \frac{k_i k_j}{2m}] \quad (7)$$

In Formula 7, A_{ij} represents any element of network adjacency matrix. If i links with j, then $A_{ij}=1$.If not, $A_{ij}=0$. $m=\frac{1}{2}\sum_{ij}A_{ij}$ represents the total margins and $k_i=\sum_i A_{ij}$ shows the degree of *i* while the number of community *i* attached is defined as O_i .When a node belongs to no more than a community, we know EQ equaling to Q. However, when all nodes belong to the same community, $EQ=0$. More obviously, the higher the value of EQ is, the more scientific the overlapping structure is.

Table 2. The Data and Parameter of the Real Network

Network	karate	dolphins	polbooks	football	jazz	email
Node	34	62	105	115	198	1133
Margin	78	160	441	613	2742	5451

Table 3. The EQ Value of the Real Network

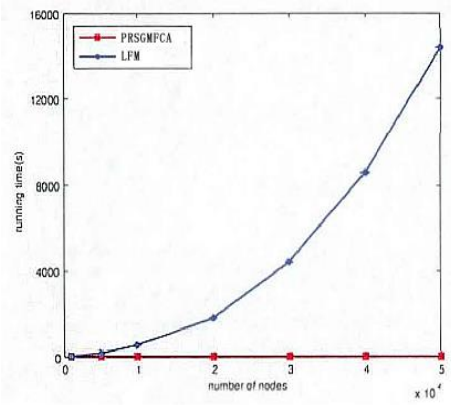
	karate	dolphins	polbooks	football	jazz	email
CPM	0.1858(k =3)	0.388(k =3)	0.5639(k =4)	0.4449(k =4)	0.2044(k =10)	0.2191(k =4)
LFM	0.3948	0.37765	0.5113	0.4399	0.2863	.....
GCE	0.2936	0.4564	0.5864	0.4867	0.2894	0.3240
PRSGMFCA	0.4025	0.4652	0.5709	0.4914	0.4276	0.3303

In Table 3 lists the results calculated by PRSGMFCA, CPM, LFM and GCE used in the real network. The EQ of PRSGMFCA is larger than that of CPM and LFM while the GCE does well in the Network football. We should know when $\gamma=1$, we can use PRSGMFCA to find 2 communities in network karate; but when EQ$(\gamma=1.2)$ is the largest value, 4 communities can be found. We almost can't find 2 communities but can find 4 communities which are calculated repeatedly in LFM, for we choose seed nodes at random. PRSGMFCA has stable results while LFM does not run enough stably to get rid of the negative effect from the wrong original node which will result in many independent single-node communities combined by many independent nodes. When we set the parameter k (the value of clique), we can get good results from CPM. However, when the network is not too dense or k is not suitable, the result will be bad. When we use the defaults, we find that the results from GCE rank between the results from PRSGMFCA and the results from LFM. But in the network football GCE performs better than PRSGMFCA while worse in the network karate. And in the network email LFM can not find correct community structure but only find a whole community and many independent

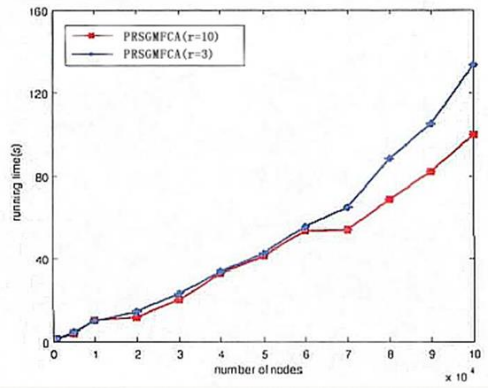
nodes, which results from the realization of algorithm. From what we have discussed above, we know that PRSGMFCA can perform stably and get ideal results as well.

C. The Complexity Analysis of Algorithm Time

After inputting the network G and the relative parameters and through some procedures, it will output a set of local community structures. During the whole process, the worst complexity of algorithm time in node centrality's ranking is $O(n^2)$, in which n represents the total nodes and in conclusion the fewer nodes, the lower complexity. Therefore, in the real world the PageRank will be weakened in the linear time $O(\log n)$, however, it will be difficult to solve the time complexity of local community expansion, for its expansion depends on the variable γ under the dynamic change of it. So in this paper it first ensures the value of γ ,then uses $O(n_c^2)$ to represent the time complexity of constructing local community with n_c nodes. The worst condition is only to find 1 whole network community as the local community, at which the time complexity is $O(n^2)$. In fact, this will not usually be found in the ground-truth network. The algorithm runs so effectively and when the community is small enough, the time will approximate to linear time. From the diagram 2, it can be observed that the algorithm generates the running time, in which the number of nodes is controlled from 1000 to 100000.



(A)The Running Time Analysis Diagram between PRSGMFCA and LFM



(B)Running Time of PRSGMFCA in Different Resolution

Figure 2. The Complexity Analysis Diagram of Algorithm Time

4. Conclusion

To solve the problems uncovered in detecting the overlapping community structure, in this paper the optimized and improved PRSGMFCA Model algorithm was used. During the process of choosing the seed node of local community, it mainly used the optimized PageRank algorithm to rank, then constructed multi-view recognition modularity based on Spin-glass model to realize local community expansion and detected the structure with the improved Greedy Search Method to obtain the clear overlapping community structure finally. Meanwhile, after analyzing the difference between the optimized PRSGMFCA and the traditional technology method in the generated and real networks, it was found that the former owns better stability, higher correct rate and a complexity of algorithm time within an acceptable range.

References

[1] M. A. Porter, J. P. Onnela and P J. Mucha, "Communities in networks [J]", Notices American Mathematical Society, vol. 56, no. 9, (**2009**), pp. 1082-1097,1164-1166.
[2] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", Nature, vol. 435, no. 7043, (**2005**), pp. 814-818.
[3] J. Yang and J. Leskovec, "Structure and Overlaps of Ground-Truth Communities in Networks [J]", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, no. 2, (**2014**), p. 26.
[4] A. Lancichinetti, S. Fortunato and J. Kertesz, "Detecting the overlapping and hierarchical community structure in complex networks [J]", New Journal of Physics, vol. 11, no. 3, (**2009**), pp. 13-15.
[5] C. Mu, Y. Liu and Y. Liu, "Two-stage algorithm using influence coefficient for detecting the hierarchical, non-overlapping and overlapping community structure [J]", Physica A: Statistical Mechanics and its Applications, vol. 408, (**2014**), pp. 47-61.
[6] L. Pan, C. Wang and J. Xie, "Detecting Link Communities based on Local Approach", In Proceedings of 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI11), (**2011**), pp. 884-886.
[7] R. K. Žalik and B. Žalik, "A local multiresolution algorithm for detecting communities of unbalanced structures [J]", Physica A: Statistical Mechanics and its Applications, (**2014**), vol. 407, pp. 380-393.
[8] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection", Physical Review E., vol. 74, no. 1, (**2006**), pp. 016110.
[9] R. Aldecoa and I. Marín, "SurpriseMe: an integrated tool for network community structure characterization using Surprise maximization [J]", Bioinformatics, vol. 30, no. 7, (**2014**), pp. 1041-1042.
[10] E. Le Martelot and C. Hankin, "Fast multi-scale detection of overlapping communities using local criteria [J]", Computing, (**2014**), pp. 1-17.
[11] H. Roitman, A. Raviv and S. Hummel, "Microcosm: visual discovery, exploration and analysis of social communities[C]", Proceedings of the companion publication of the 19th international conference on Intelligent User Interfaces, ACM, (**2014**), pp. 5-8.
[12] S. Fortunato and M. Barthelemy, "Resolution limit in community detection", Proceedings of the National Academy of Sciences, vol. 104, no. 1, (**2007**), pp. 36-41.
[13] W. W. Zachary, "An Information Flow Model for Conict and Fission in Small Groups", Anthropological Research, vol. 33, (**1977**), pp. 452-473.
[14] A. Alamsyah and B. Rahardjo, "Community Detection Methods in Social Network Analysis [J]", Advanced Science Letters, vol. 20, no. 1, (**2014**), pp. 250-253.
[15] M. E. J. Newman, "Modularity and Community Structure in Networks", Proceedings of the National Academy of Sciences, vol. 103, (**2006**), pp. 8577-8582.
[16] B. Liu and T. Qian, "A Local Greedy Search Method for Detecting Community Structure in Weighted Social Networks [M]", Advanced Data Mining and Applications. Springer Berlin Heidelberg, (**2013**), pp. 360-371.
[17] H. Shen, X. Cheng, K. Cai and M. B. Hu, "Detect overlapping and hierarchical community structure in networks", Physica A: Statistical Mechanics and its Applications, vol. 388, no. 8, (**2009**), pp. 1706-1712.
[18] Y. Geng, J. He and K. Pahlavan, "Modeling the Effect of Human Body on TOA Based Indoor Human Tracking [J]", International Journal of Wireless Information Networks, vol. 4, (**2014**), pp. 306-317.

Author

Zhang Xin (1984-), male. His main research area is data mining.

