International Journal of Grid and Distributed Computing
Vol. 9, No. 12 (2016), pp.115-126
http://dx.doi.org/10.14257/ijgdc.2016.9.12.11

Community Discovering Algorithm based on Nodes with Maximum Degree and Label Propagation

Li Shengli[1], Chen Deyun[2] and Yao Yuanzhe[3]

[1,2]*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China*
[3]*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China*
lisl@hrbust.edu.cn, chendeyun@hrbust.edu.cn, yzyao@tsinghua.edu.cn

Abstract

Label propagation algorithm (LPA algorithm) is gotten widely attention by its simplicity, rapidity and greater effectiveness. Aimed at the problem of label updating process is sensitive to the order of nodes in LPA algorithm, an improved algorithm is proposed in this paper. The algorithm starts from the nodes with maximum degree in the network, according to its neighbor nodes' community similarity index to decide whether its label setting, and then completes label setting in the first round; On this basis, it continues to carry out the iteration in label propagation algorithm, completes the entire network of community structure detection. The experimental results show that the improved algorithm is slightly better than the original algorithm at running time and number of iterations, has higher robustness, prevents the occurrence of the trivial solution.

Keywords*: Community Detection, Label propagation algorithm, Nodes with Maximum degree, Neighbour Node*

1. Introduction

Complex network research is in a vigorous development stage[1,3], the idea has filled every corner of the science and society. Complex networks is a abstraction of complex system, in reality many complex systems can be described by the related properties of complex networks and be analyzed. Such as the Internet, WWW[1], social networking[1,2], biological network[4], electricity and transportation network[6], economic and financial network[1], *etc*.

Communities are the common characteristics of complex network structure, there are close relationship between the function of complex networks and its community structure, such as robustness, transitivity, *etc*.), hence discovering a correct network community structure is of great significance to study its related properties[1,2,3,4,5,6]. In recent years, the discovery and analysis of the community structure of complex networks have the attention of many researchers, many excellent community discovery algorithm have emerged in large numbers. However, because many systems are too large, the time complexity and accuracy are still the two main problems in the analysis algorithm of large-scale complex network community structure[1,2,5]. Therefore, how to find a fast and reliable network community structure analysis algorithm according to the characteristics of different types of networks is still the main problem to be solved in the future.

At present, there are a lot of concerns about discovering the global community structure in the network, but when the network is too large or too quickly evolution, most of the methods that require the global knowledge of the network will become unrealistic[7,18]. However, many times we only focus on the community structure of few

ISSN: 2005-4262 IJGDC

given nodes, rather than all nodes. So it is more appropriate to find and focus on the local community structure of the network[7,11,12,13].

Researchers have proposed the community discovery method based on local information network [7,18,19,20]. In literature [7], the method of finding community structure by using the core node instead of using the initial node to calculate the local module is proposed. The method is not sensitive to the location of the source node. Literature [20] proposed a community discovery algorithm based on searching neighbor nodes, the algorithm starts from the nodes with maximum degree score in the network, combines with the neighbor nodes searching and neighbor nodes voting, then expand the search from the local of the network to the whole, and finally forms multiple disjoint communities, the running time of the algorithm is close to O(m+n), and the accuracy of corresponding community structure obtained by the algorithm is higher.

Raghavan et.al[11] proposed a fast label propagation algorithm, which avoids the restriction brought by the prior information, algorithm in a simple, rapid, effective and widely attention[15,17]. However, subsequent research[12,13,14,15] has shown that LPA labels are sensitive to the node order update process, the different update order tend to have different corporate results, and trivial solution easily comes into being, namely all the nodes in the same community, and thus lose the significance of community partitioned.

This paper proposes an improved algorithm against LPA algorithm, which is very sensitive to the order of nodes, combined with the community discovery algorithm based on neighbor node searching. The algorithm begins from the nodes with maximum degree in the network, according to the community similarity index of the neighbor nodes, sets their labels for the neighbors that belong to the community include them, thus completes the LPA algorithm in the first round of the label setting; It Continues to carry out label propagation iteration, and then quickly completes the whole network community structure discovery.

2.Basic & Related Works

2.1. LPA Algorithm and its Improvements

In many cases, the edge of a complex network is a representation of the information spread among individuals, and the result is usually the same information that is shared among the nodes in the community. Raghavan *et al*. [11] Proposed a fast label propagation algorithm (LPA algorithm) based on this idea. As shown in figure 1, LPA algorithm firstly assigned to each node unique label, every step in the iteration, each node will update its label from the label occurring with the highest frequency among its' neighbors, if there are more than, it will randomly select a label as an update from these labels, after several iterations densely connected nodes will converge to the same label, Eventually the nodes with the same label are classified as a community.

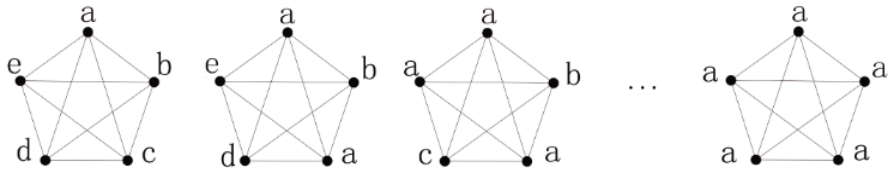


Figure 1. The Process of Label Propagation

The implementation steps of label propagation algorithm are as follows[11]:

1. Initialize the labels of all nodes in the network, for a given node x, Cx(0)=x.

2. Set t=1.

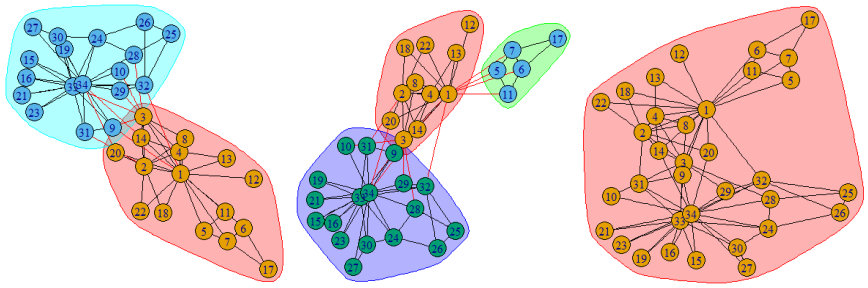
3. Arrange the nodes in the network in a random sequence and set it to X.

4 . For each $x \in X$ chosen in that specific order, let $Cx(t)=f(Ci1(t),\ldots,Cim(t),Ci(m+1)(t-1),\ldots,Cik(t-1))$. f here returns the label occurring with the highest frequency among neighbors and ties are broken uniformly randomly.

5. If every node has a label that the maximum number of their neighbors have, then stop the algorithm. Else set t=t+1 and go to 3.

6. Return C.

The biggest advantage of the LPA algorithm that does not require any input parameters, such as the community of the number, size, *etc*.. Moreover, the algorithm has linear time complexity degree with O(m+n), m is the number of edges in the network, n is the number of nodes in the network, its convergence speed is very fast. Usually by 5 iterations, 95% of nodes or more are classified correctly[11]. It can beapplied to large scale complex network. However, this method is faster but produces different results in each run based on initial configuration. So one need to run the algorithm several times and built the consensus, but it consumes time[12,13].



a. Modularity with 0.3714661, b. Modularity with 0.3990796, c. Modularity with 0

Figure 2. The Results Obtained by using the LPA Algorithm in the iGraph Program to Deal with the Zachary Network Many Times (More than one, here only 3)

Barber *et al*. improved the label updating rules, respectively formed LPAm[12]、LPAm+[14]、LHLC[13]、MLPA[15]and other improved algorithm. Barber [13] and Liu[14] *et al*. improved updating formula of label in LPA, maximizing the modularity measuring values, the new algorithm is shorthand for LPAm. Based on LPA algorithm, using the structure similarity calculation, through the biggest constraint propagation model, MLPA updates the node's label, makes the results of community partition more in line with the corporate internal structure is relatively close, associations between structure is relatively sparse features and improves accuracy of community partition.

2.2. Community Discovering Methods based on the Nodes with Maximum Degree

Firstly, the value of a node in the network depends on the location of the nodes in the network, the value of this position the centrality of the nodes. In network analysis, degree centricity is the most direct measure, namely, the greater the degree of a node means that the node is more important [1,2].

Previous studies have shown that for certain problems, the node with maximum degree is not important. A generally accepted view, that is the importance of a node depends on the number of neighbors (degree score) also depends on the quality of the neighbors [2,8].

Through the study the community division results of the GN algorithm, Zhang *et al*. [7] concluded that the inner 98% nodes has a lower degree than their adjacent nodes, and the nodes with the core node has a higher degree.

There are some key nodes in the real network, these key nodes are usually the leader of the association, which are connected with many nodes in the network, which have the characteristics of large degree [18,20]. So, First we can find out the node set with higher degree in the network, and then select from these nodes as the core of communities, search its neighbor nodes and calculate the community similarity, respectively join the eligible nodes to the communities which take these core nodes as the core.

Therefore, due to the complexity of the network, we can find the approximation "core" or the node closer to the "core", rather than the "absolute" core node in the network; And then combine the relevant methods to complete the community division of the whole network. In the discovery of the core node, we can follow the following rules:

（1）The number of nodes with maximum degree (greater than or equal to 3) is 1 and these nodes are not adjacent to the existing core node, which is the core node;

（2）The number of nodes with maximum degree(greater than or equal to 3) is more and these nodes are adjacent, we can select the node as the core of community, which is not adjacent to the existing core nodes and has larger clustering coefficient (the node with the smaller clustering coefficient generally has the external connection with outer-community)；If there are more similar nodes (adjacent and with the same number of degrees and clustering coefficients) we can choose any one of them as the core node;

（3）The above nodes are adjacent to the existing core nodes, but there is the bridge connection between them, such as the connections between the nodes 7 and 8 in Figure 3 is the bridge connection, they are not in the same community.

The nodes conform to the "core node" conditions in real network, such as node 34 and 1 in Zachary network; node 9, 7, 2 in figure 3.

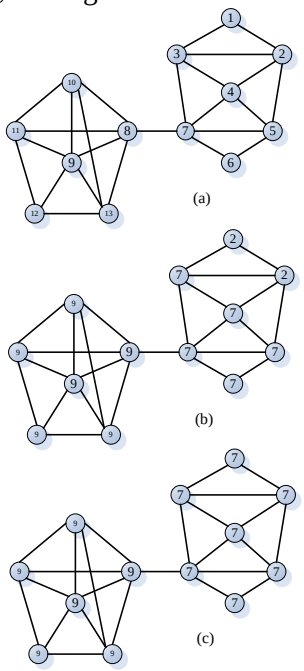


Figure 3. The Process of Improved Algorithm

In figure 3, Sample network with 13 nodes and 25 edges. The degree of node 9,8,7 is 5, and its clustering coefficient is C9=0.7, C8=0.5, C7=0.3., respectively. According to the above rules, we take the initial decision nodes 9,7 and 2 as the "core node". (a)primal graph；(b)The first round of the label setting, according to the core node to classify

community; (c)Continue to perform the iteration in LPA algorithm, the network is divided into two communities.

2.3. Index of Two Neighbors Nodes' Community Similarity based on Local Information

The network $G = (V, E)$ contains n nodes, and its adjacency matrix is A. If there is edge between the node V_i and the node V_j, then $A_{ij} = 1$, else $A_{ij} = 0$. For the node pairs (V_i, V_j), if there is edge between them, then the node V_i is the neighbor node of the node V_j, at the same time, the node V_j is the neighbor node of the node V_i.

2.3.1. Index of Neighbors Nodes' Community Similarity based on Local Information

In the social network, friends of friends also is likely to be new friends between them. For general network, we can be promoted that the more common neighbor nodes of the two nodes, the two nodes are more similar, which tend to be connected to each other. Therefore, the node similarity index based on common neighbors can be defined.

If all the neighbor nodes of the node V_i is $N_i = (V_{i1}, V_{i2}, \dots, V_{ik},)$, all the neighbor nodes of the node V_j is $N_j = (V_{j1}, V_{j2}, \dots, V_{jk},)$,

$$S_{ij} = \left|N_i \cap N_j\right| \left(d_j > 1\right) \quad (1)$$

Here, $|x|$ means the element number of x.

2.3.2. Index of Two Neighbors Nodes' Community Similarity based on Local Information

When the number of the same elements in N_j and N_i is greater than a certain amount, it is considered that the node V_i and $V_j$belong to a community. So the node similarity index of two neighbors' nodes based on common neighbors can be defined.

$$\beta_j = \frac{\left|N_i \cap N_j\right|}{d_j - 1} \left(d_j > 1\right)$$

$$\beta_j = 1 \left(d_j = 1\right) \quad (2)$$

Here, d_j is the degree of node V_j, ɣ is a index.

For two neighbor node V_i and V_j, if V_i is the core node and $\beta_j \geq$ ɣ, then V_j belongs to the community with the core node V_i. When $d_j = 1$, it shows that V_j is the least important nodes in the network, if partitioning community, We can assign it to the only community where its neighbor nodes lies.

3. Community Discovering Algorithm Based on Nodes with Maximum Degree and Label Propagation

3.1. Algorithm Introduction

The algorithm starts from the nodes with maximum degree in the network, according to its neighbour nodes' community similarity index to decide whether its label setting, and then completes label setting in the first round; On this basis, it continues to carry out the iteration in label propagation algorithm, completes the entire network of community structure detection. The algorithm process is simply shown in Figure 3.

3.2. Algorithm Framework

Algorithm 1 An Improved Community Discovering Algorithm Based on Nodes with Maximum degree and Label propagation

Input: Graph $G = (V, E)$

Output: Community Structure C

1. Initialize the labels of all nodes in G, for a given node x, Cx(0)=x, and set them to no-visited status.

2. Starts the first round label setting, for all nodes which degree is greater than 3, find the maximum number of degrees maxdeg=max(degree(V)).

3. For the nodes which degree number equal maxdeg and have not been visited, setting the degree of vi to 0.

4. For the node vj which is vi's neighbor and not been visited, calculates the community similarity between them βj, if $\beta j > \gamma$, then uses the label of vi to set the label of vj, set the degree of the node vj to 0, at the same time, the node vj is marked as the visited.

5. The end of the first round label setting.

6. continues to LPA iteration to set the label of the node

7. the end of LPA iteration

8. the processing of tiny community, take the community whose number of nodes is less than 3 into the nearest community

9. return C.

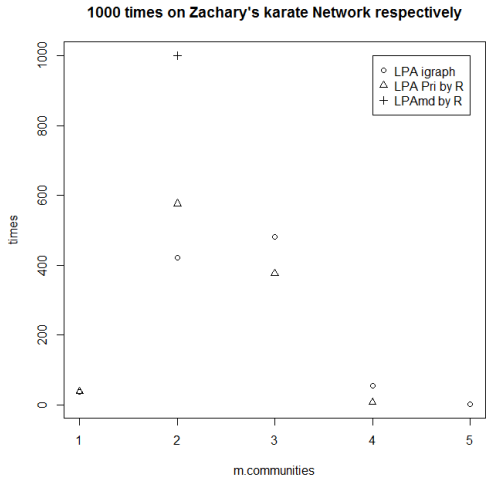
4. Experimental Results and Analysis

4.1. Experimental Results

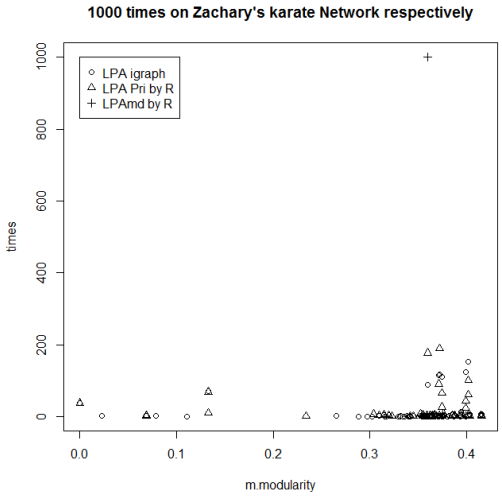
All these experiments are conducted on a machine with i5-3470 processor, 4GB primary memory space and 500GB secondary memory using RGui (64-bit)[20] software with version 3.2.3(2015-12-10). The community detection algorithms used in this paper are implemented in igraph package[21] and the R scripts are developed by utilizing this package.

In these experiments, we use R language to achieve the LPA algorithm and improved LPA algorithm based on the core node, and compare the operation results with the LPA function in the iGraph package. For convenience of expression, respectively referred to as LPA Pri by R, LPAmd by R, LPA igraph.

A. Zachary Karate Club, this example is built on Zachary Karate Club, which presents data from a university-based karate club for n = 34 vertices, in which a factional division led to a formal separation of the club into two organizations. A disagreement developed which resulted in instructor leaving and starting a new club, taking about a half of the original club members. Fig. 2(a) show the original karate club。

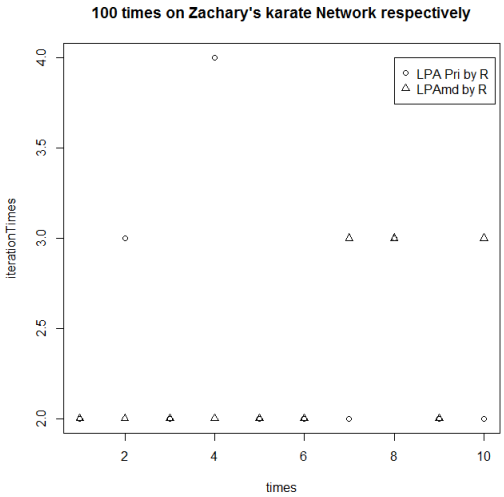


(a) distribution of the communities number obtained by the program running 1000 times respectively



(b) distribution of the modularity values obtained by the program running 1000 times respectively

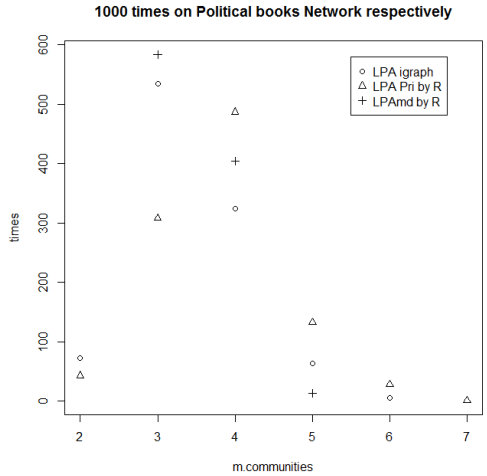
It can be seen that the results of LPA Pri by R and LPA iGraph are basically consistent, relatively dispersed, there are trivial solutions. The results of R. LPAmd by R are more stable, the network classified into two groups, It basically conforms to the actual grouping of the network.



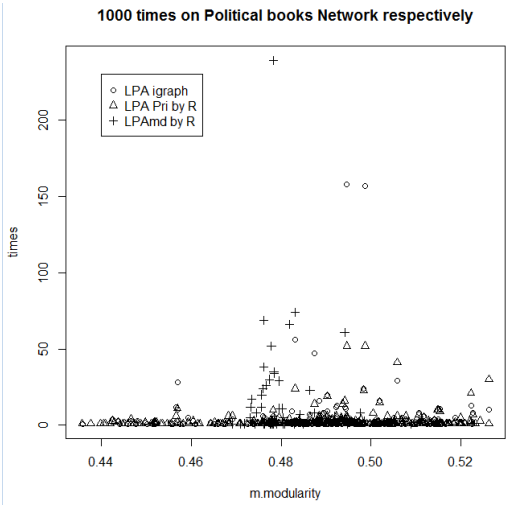
(c) Contrast figure of iteration times obtained by the program running 100 times respectively

Figure 4. The Community Detection Statistical Results based on Zachary Karate Club Network

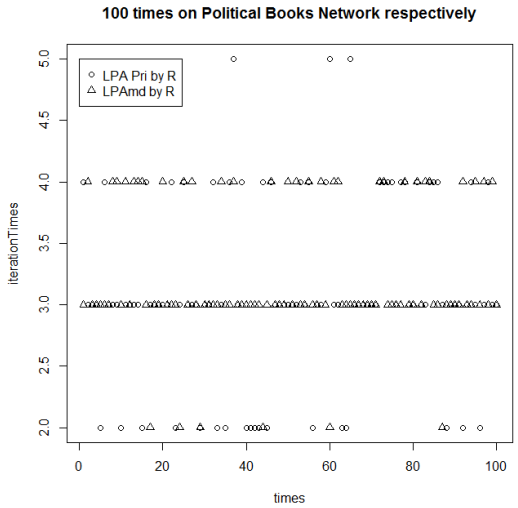
B. Political books, this data is the Books about US politics. Nodes represent books about US politics sold by the online bookseller Amazon.com. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books' feature on Amazon .Nodes have been given values "l", "n", or "c" to indicate whether they are "liberal", "neutral", or "conservative". This network has 104 nodes and 441 edges. And this network can be partitioned into three community factually.



(a) distribution of the communities number obtained by the program running 1000 times respectively



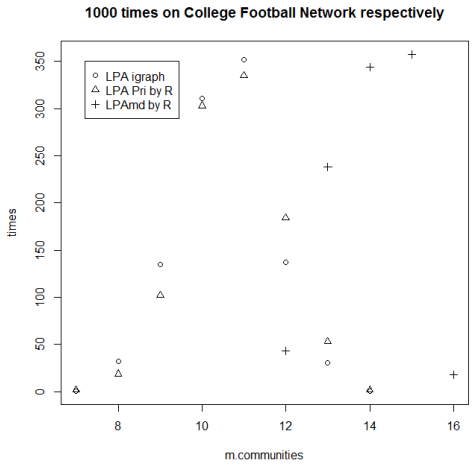
(b) distribution of the modularity values obtained by the program running 1000 times respectively



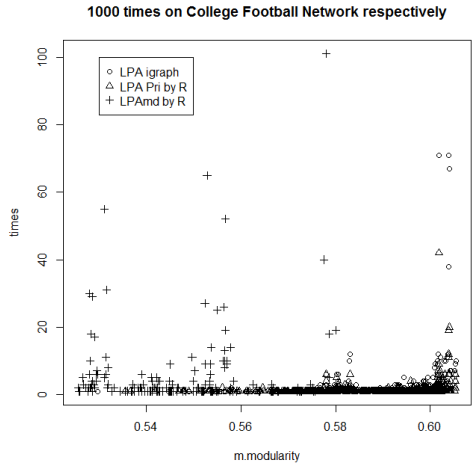
(c) Contrast figure of iteration times obtained by the program running 100 times respectively

Figure 5. The Community Detection Statistical Results based on Political Books Network

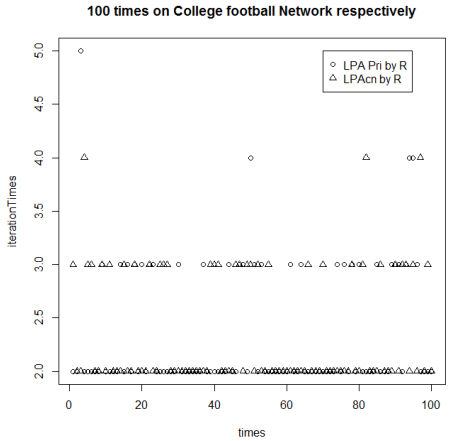
C. American College football, a network of American football games between Division IA colleges during regular season Fall 2000. In the network, each node represents the attend team football season in 2000 colleges and universities in the United States, the connection between two nodes represents the edge of the corresponding at least between the two teams had a race.



(a) distribution of the communities number obtained by the program running 1000 times respectively



(b) distribution of the modularity values obtained by the program running 1000 times respectively



(c) Contrast figure of iteration times obtained by the program running 100 times respectively

Figure 6. The Community Detection Statistical Results based on American College Football Network

From the above experimental results, we can conclude that the improved algorithm is more stable than the original algorithm, and the number of communities and the modularity value is more concentrated, and the number of iterations is less than the

original algorithm. According to the experimental results for different networks, we found some shortcomings of the improved algorithm.

（1）This method is suitable for the network with wide connections to center nodes, not suitable for a star-shaped on macro, but less connected to the center of the network. . Through the analysis of these networks,we can divide the network into multiple sub networks combining with the node or edge betweenness,

（2）This method has strong randomness.There are many partition results , it usually needs to combine the network background to judge the appropriate result artificially.

4.2. Algorithm Analysis and Optimization

Compared with the original algorithm, in the first iteration, the improved algorithm computes the degree of each node and it needs O(n) time. Then for each node with degree greater than 3, we discover its neighbor nodes and determine the community of these nodes, it takes O(dx) time, all processing time is still O(m). In later iterations, the original algorithm is still used, so the time of the whole algorithm is still O(n+m).

In the process of experiment, if the label occurring with the highest frequency among neighbors, the determined node's label is set to this label; But if there are multiple nodes with highest label, randomly, it would increase the instability of the results, so this need to be explicitly set. There are two kinds of processing methods: (1) According to the idea of priority in the label propagation of the node with greater degree, we can compute the sum of the corresponding node degrees in these labels, and choose the label of the nodes with maximum number. (2) According to the idea of the priority of small groups,we choose the label of the nodes with minimum number.. In the implementation of our algorithm, we choose the first method.

5. Conclusion and the Future Works

In this paper, we first focus on some given node in the network, according to the local information of the network. In order to solve the problems of LPA algorithm which is very sensitive to the order of nodes,we propose a community discovery algorithm, which starts from the node with the maximum degree, according to the community similarity index of the neighbor nodes, sets their labels for the neighbors that belong to the community include them, thus completing the the first round of the label setting of LPA algorithm; continues to run label propagation and rapidly completes the whole community structure discovery in the network.

In the later study, we will still exert the advantages of low time complexity of this method, aim at the problem of stronger randomicity, using these methods and technologies such as modularity, information entropy to improve the algorithm, in order to reach a relatively short time to get more accurate community division.

In the future, parallel computing will be one important aspect of our research. In small-scale applications, we can make full use of GPU's ability, send the part of data processing tasks to the database system, realize community partition algorithm of small parallel processing. On the processing of large data sets, we can upload huge amounts of data to the HDFS, use HADOOP to analyse HDFS data, realize large network community partition massively parallel processing with massive data.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61370202).

References

[1] X.-F. Wang, X. Li and G.-R. Chen, "Network Science: An Introduction. Higher Education Press", Beijing, **(2012).**
[2] L. Tang, H. Liu, Y.-M. Wen and Y.-Z. Bi, "Community Detection and Mining in Social Media", China Machine Press, Beijing, **(2013)**.
[3] M. Newman, URL: http://www-personal.umich.edu/~mejn/, **(2015)**.
[4] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Natl. Acad. Sci., USA, **(2002)**, pp. 7821-7826.
[5] A. Clauset, M. E. J. Newman and C. Moore, "Finding community structure in very large networks", Physical Review E, vol. 70, **(2004)**.
[6] S. R. Chintalapudi and M. H. M. K. Prasad, "A Survey on Community Detection Algorithms in Large Scale Real World Networks", 2nd International Conference on Computing for Sustainable Global Development (INDIACom), **(2015)**.
[7] T.-T. Zhang and B. Wu, "A Method for Local Community Detection by Finding Core Nodes", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, **(2012)**.
[8] M. Kitsak, L. K. Gallos and S. Havlin, "Identifying influential spreaders in complex networks", Nature Physics, **(2010)**.
[9] P.-F. Wang, "Research of Community Detection Algorithm base on Node Degree Difference and Node Similarity", Lanzhou University, **(2014)**.
[10] J.-G. Liu, Z.-M. Ren, Q. Guo and B.-H. Wang, "Node importance ranking of complex networks", Acta Physica Sinica, **(2013)**.
[11] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large scale networks", Phys Rev E, **(2007)**.
[12] M. J. Barber and J. W. Clark, "Detecting network communities by propagating labels under constraints", Phys Rev E, **(2009)**.
[13] I. X. Y. Leung, P. Hui and P. Liò, "Towards real-time community detection in large networks", Phys Rev E, **(2009)**.
[14] X. Liu and T. Murata, "Advanced modularity-specialized label propagation algorithm for detecting communities in networks", Physica A, **(2010)**, vol. 389, no. 7, **(2010),** pp. 1493–1500.
[15] J.-B. Huang, X. Zhong, H.-L. Sun and W.-T. Mao, "A Network Community Detection Algorithm via Constrained Label Propagation with Maximization of Similarity-Based Modularity", Acta Scientiarum Naturalium Universitatis Pekinensis, Beijing, vol. 10, **(2012)**, pp. 388-396.
[16] S. Gregory, "Finding overlapping communities in networks by label propagation", New Journal of Physics, vol. 12, no. 10, **(2010)**, pp. 103018.
[17] Y. Wang, "Researches of Community Detection Algorithms in Complex Networks", Anhui University, Hefei, **(2014)**.
[18] S. Zhao, M.-Z. Wu, Z. Duan, Y. Wang, Y.-P. Zhang, "Community Detection Algorithm Based on Neighbor Searching", Journal of Chinese Computer Systems, vol. 08, **(2015)**, pp. 1795-1798.
[19] D.-W. Zhang, F.-D. Xie and Y. Zhang, "Fuzzy analysis of community detection in complex networks", Physical A, vol. 389, **(2010)**, pp. 5319-5327.
[20] R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, URL: https://www.R-project.org/, **(2015)**.
[21] igraph – The network analysis package, URL: http://www.igraph.org/, **(2016)**.
[22] A. Lim, W. Tjhi, L.-Y. Tang, "R High Performance Programming", Publishing House of Electronics Industry, Beijing, **(2015)**.

