A Similarity Based Community Division Algorithm

Lingjuan Li, Wei Wang
College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China
E-mail: fqlilj@163.com

Abstract—**Community division is an important research topic in complex network area. In order to quickly and accurately find community structures in complex network, a similarity based community division algorithm named SCDA is proposed in this paper. This algorithm finds community structures by clustering the nodes according to the importance of nodes and their similarities. It selects the node owning greater clustering coefficient as the cluster center, puts the nodes with similarity larger than given threshold to the cluster, and iterates the process until the node collection is empty. Then the clusters generated by the algorithm are communities. SCDA reduces the time complexity by starting from the important node and ignoring those nodes having been clustered when determing new cluster center, and it promotes the division accuracy by using clustering coefficient and node similarity properly. The experimental result of applying the algorithm to the classical social network, the Zachary's Network of Karate Club Members, shows that SCDA costs less time and has higer accuracy. The algorithm SCDA is valid in community division.**

Keywords- complex network; community division; clustering; similarity;

I. Introduction

Complex network is a network topology map composed of a lot of nodes and perplexing relations between them. In real life, complex network can describe social relations, such as the communication between people, the relationship between paper collaborators, the relationship of food chain between species, etc. With the indepth study of the physical meaning and mathematical characteristics of the network, it was discovered that many complex networks have a common nature, i.e. community structure. Revealing the community structure in the network has very important significance for understanding the network structure and analyzing network performance.

A common definition of network community is as follows: community is a collection of vertices that are densely connected between themselves while being loosely connected to the rest of the network [1].

At present, the research on community division is mainly focused on two types of methods: method based on optimization and heuristic method.

Methods based on the optimization divide the community structure of complex networks by optimizing the predefined target function, and the method based on the modularity function Q always tends to find a rough rather than a fine network community structure. The main representative algorithms are spectral bisection algorithm [2] and Kernighan-Lin algorithm [3].

Heuristic methods divide community structure in complex network by designing heuristic rules. And the feature is to design the heuristic algorithm based on some intuitive assumptions. For most of the social network, they can quickly find the optimal or approximate optimal community structure. The main algorithms are GN（Girvan-Newman）algorithm [4] and Newman greedy algorithm [5].

All of above algorithms have their own advantages and disadvantages. The spectral bisection method has rigorous mathematical theory, and can divide quickly, but for the social network without very obvious community structure, the division result is not accurate. The Kernighan-Lin algorithm is a greedy algorithm, and it is necessary to know the number of community in advance. GN algorithm detects community structure according to the edge betweenness from top to down, and needs recalculate the edge betweenness in each dividing process, so it will cost more time.

Based on the idea of heuristic algorithm, this paper proposes a similarity based community division algorithm SCDA. The algorithm selects the node owning greater clustering coefficient as the cluster center, puts the node with similarity greater than given threshold to the cluster, and iterating the process until the node collection is empty.

II. Design of SCDA

A. Related Definitions

In order to facilitate the study of community structure, the researchers have givn some quantitative definitions about the community, such as node degree, clustering coefficient, etc. In this paper we adopt these definitions, and give the definition of node similarity. The node similarity is the important basis of clustering in SCDA. The definitions of node degree, clustering coefficient and node similarity are as follows:

Node degree is defined as the number of other node which has edge connected to the node.

Clustering coefficient C_i is defined as: $C_i=E_i/T_i$, $C_i\in[0,1]$. If node degree of node i is k, then E_i is the actual number of edges between k neighbor nodes of node i, T_i is the maximum possible number of edges between k neighbor nodes of node i, $T_i=k(k-1)/2$. When $C_i = 1$, node i is in a central position.

In this paper, node similarity is quantitatively calculated by Euclidean distance based on the idea that more similar the shortest path of two nodes to other nodes, more similar these two nodes.

If G is an undirected graph owing M nodes, G = <V, E>, V={x_1, x_2, x_3, …, x_m}, x_{ik} is the shortest path between node i and node k, and x_{jk} is the shortest path between node j and node k. Euclidean distance of node i, j is defined as:

$$D(x_i, x_j) \equiv \sqrt{\sum_{k=1}^{M} \left(x_{ik} - x_{jk}\right)^2} \qquad (1)$$

Since there is an inverse relationship between the Euclidean distance and similarity, to ensure the similarity in the range (0, 1], we define node similarity between node i and node j as:

$$S(x_i, x_j) = \frac{1}{1 + D(x_i, x_j)} \qquad (2)$$

B. *The Steps of SCDA*

Algorithm: SCDA

Input: An undirected and unweighted network G = <V, E>, V is the set of nodes, E is the set of edges.

Output: Community structures

Steps:

- Calculate clustering coefficient of all nodes and store results in a key-value pairs in the map, key is the node number and value is the clustering coefficient.
- According to the (2), calculate the similarity between all nodes and store results into a similarity matrix.
- From the set of clustering coefficient, select the node owning the largest clustering coefficient as a cluster center.
- From the similarity matrix, use the average similarity value of all nodes as a threshold value St. Then put the node with similarity greater than St to the cluster and delete those nodes from the node set.
- Determine whether the node set is empty, if it is empty, stop. Otherwise, go back to the third step.

III. Analysis of SCDA

A. *Comparing with the Kernighan-Lin and Spectral Bisection Algorithm*

Kernighan-Lin algorithm is only applicable to complex network owning two community structures, whereas the SCDA algorithm does not need to know the number of community, i.e. it can automatically identify the number of communities. Spectral bisection algorithm requires community structure to be obvoious. On the contrary, the SCDA algorithm has not such requiment.

B. *Time Complexity of SCDA*

The SCDA algorithm needs to calculate the similarity between any two nodes. If there are n nodes, the time complexity of this part is O (n^2). Both the third step and the fourth step of SCDA algorithm require n^2 times comparison, so the time complexity of this part is also O (n^2). Therefore, the average time complexity of SCDA is less than that of spectral bisection and GN algorithm, O (n^3).

IV. Test of SCDA

In order to verify the feasibility and accuracy of the SCDA algorithm, we choose the classical social network Zachary's Network of Karate Club Members [6] to do division.

A. *Zachary's Network of Karate Club Members*

Zachary's karate club, as a karate club in an American university, is a social network which has 34 nodes and 78 edges. Each node represents a club member, and each edge means that there's social interaction between two club members. This club split into two independent clubs due to internal divergence. One club was led by the original coach and the other was led by original director. Fig.1 shows the network and its community structures.

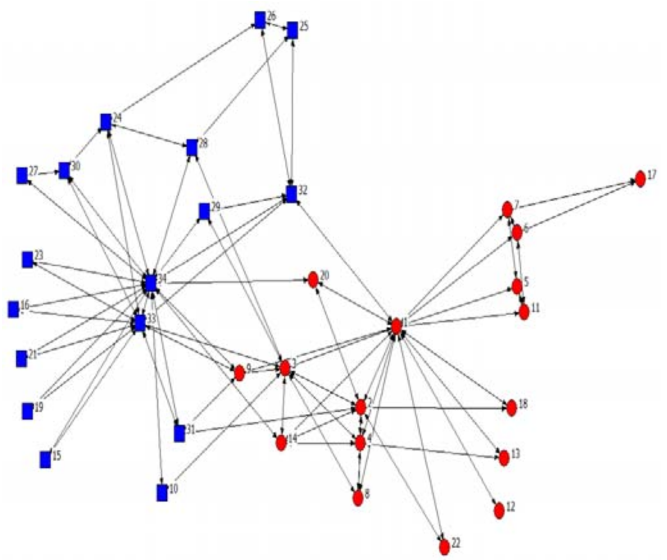

Fig. 1. Community structure in Karate Club network.

B. *Experimental Result*

We selects Pentium T4200 processor, 2G memory and windows 7 operating system as the operating platform, and the algorithm is implemented in java and running on myeclipse. The experimental result is shown in Table I. And the running time is 22ms.

TABLE I. Experimental Result

Community Center	Nodes in the Community
Node 34	9,15,16,19,21,23,24,25,26,27,28, 29,30,31,32,33,34
Node 1	1,2,3,4,5,6,7,8,9,10,11,12,13,14 17,18,20,22,29

As can be seen from Table I, the community division result of the algorithm SCDA is basically consistent with the actual network. Only the node 9 and node 29 are repeatedly divided. The accuracy rate of the division is 94.12%.

It is noteworthy that both the node 9 and the node 29 are at the boundary of the two communities in Fig.1.

V. CONCLUSIONS

There are many community division algorithms for complex networks. This paper proposes a similarity based community division algorithm SCDA. It starts from the important node and divides community based on the similarity between the nodes. It needs not know the number of communities and does not require the community structure to be obvious, its division accuracy is better and its time complexity is lower. And the division result on Zachary karate club network has shown the advantages of SCDA.

However SCDA has a disadvantage. That is, it may repeatedly divide the boundary node to more than one community. This needs overcoming in our further study.

REFERENCES

[1] MEJ Newman, “Modularity and communities structures in networks,” Proc. of the National Academy of Science, 2006, vol. 103(23), pp.8577-8582.

[2] M. Fiedler, “Algebraic Connectivity of graphs,” Czech Math J, 1973, vol. 23, pp.298-305

[3] B.W.Kernighan, S.Lin, “An efficient heuristic procedure for partitioning graphs,” Bell System Technical Journal, 1970, vol. 49, pp.291-307.

[4] M. Girvan, MEJ Newman, “Community structure in social and biological networks”. Proc. Natl Acadthe Sci.USA . 2002, vol.99, pp.7821-7826.

[5] MEJ Newman, “Fast algorithm for detecting community structure in networks,” Proc. Natl Acad Sci. , 2001,vol. 99, pp.7821-7826.

[6] Darong Lai, “Complex network community structure analysis method research,” Shanghai Jiao Tong university, 2011.