ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

Parallel Algorithm for Identifying Overlapping Communities from Local Extension

Hua Long[1] and Baoan Li[1, 2, *]

[1]Computer School, Beijing Information Science and Technology University, Beijing 100101, China

[2]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China

*Corresponding author

***Abstract*—The overlapping communities identification in complex networks was studied with the parallel algorithm from local extension. The algorithm thought originated from clustering of modularity. Firstly, based on the Map/Reduce framework, the core node was selected according to the distribution and aggregation of nodes; Then, the algorithm calculated the members of newly added nodes and the existing small community formed by core node and neighbor nodes, and achieved a gradual expansion from local community to global community; Finally, the overlapping degree and modularity increment were used to effectively identify community. Experimental results show that: Community identification algorithm realized an extension from local community, and improved the effect of aggregation, and its parallel computing is faster and more efficient than traditional algorithms.**

Keywords-overlapping community; clustering coefficient; core node; parallel computing; overlap degree; modularty increment

I. INTRODUCTION

In the era of explosive growth of large data, the life of people is more digital and networked. For example, interpersonal network is formed between people and trading network is formed by buying goods between customers and so on. These networks are called complex network because them have a high complexity [1]. In the study of network life, people usually use a network model to represent a complex network. In the network model, nodes in the graph describe the concrete entities in life and connections in the graph describe the relationship between entities [2].

With the continuous research, it is found that there is a common phenomenon in many complex networks. The phenomenon is reflected that internal connections are close and external connections are sparse in community. In the exploration of network community, the research results of Newman and Girvan have become a central issue and research direction [3]. In depth study of the characteristics in network community, it is a great application prospect and meaningful challenge to refine and cohesion the property of nodes and edges for understanding the specific function and relevant disciplinarian in network community.

At present, there are many excellent community discovery algorithms [4-8], but most are applied in the traditional non-overlapping network community, in other words, the network is divided into several independent communities so that each node belongs and only belongs to one community [9]. However, in the actual network, a node maybe belongs to more than one community, such as; a micro-blog user is concerned about the different groups in social network. Therefore, the identification of overlapping community has become a hot research topic in recent years. In depth study, the researchers find that the overlapping is an obvious characteristic, so the quality of identifying community will reduce with ignoring overlapping. However, overlapping is difficult to be computed in communities, and it brings new difficulties to the community identification algorithms [10-14]. At the same time, the running time of traditional community detection algorithms in complex large-scale network community is slow and its efficiency is low. Therefore, the research proposes a parallel algorithm for identifying overlapping communities from local extension. The contribution in research is that core node is sought by graph theory and network nature and was applied in parallel computing. A new method of updating modularity increment and merging overlapping communities was proposed with the idea of modularity and overlapping. The advantages of algorithm are verified by comparing the running times between parallel computing and other algorithms in experiment.

II. OVERLAPPING COMMUNITY IDENTIFICATION ALGORITHM BASED ON LOCAL EXTENSION

In the actual complex network, the community begins to be formed a local small group, then it extends continuously to form total network. Therefore, small groups and local communities are the natural characteristics of community [15].

A. Core Node

The complex network community is abstracted as a network model, so it is in line with the features of network aggregation. Based on the features, the method seeks core node in community by employing point clustering coefficient and edge clustering coefficient.

1) Point clustering coefficient and edge clustering coefficient

In complex networks, point clustering coefficient usually is described the aggregation of community [16]. The point clustering coefficient is expressed as a ratio of the actual number of edges between nodes and the total number of edges that may exist.

Edge clustering coefficient represents a ratio of the parameters of closed loop with three edges. In other words, it is a ratio of the number of triangular rings that consists an edge of two end nodes and other edges of the common adjacent nodes and the maximum number of triangles that may contain the edges.

2) Core node selection

According to the above concepts of point clustering coefficient and edge clustering coefficient, the core node is found. The formula is as follows:

$$Center(i) = c(i) + \frac{\sum_{j \in N} C(ij)}{K_i} \tag{1}$$

The *c(i)* and *C(ij)* respectively represent point clustering coefficient and edge clustering coefficient. K_i represents the degree of node *i*. The core node and its neighbor nodes may be merged into a small community because they connect closely. Therefore, in a complex network, there are many core nodes and its corresponding neighbor nodes, which may be divided into a number of small communities which were extended to global communities.

B. Membership Degree of Node and Community

A complex network community is represented by $G=(V,E)$, the V and E are the collection of nodes and edges. Assuming the community G is divided into l communities including $C_1,C_2...C_l$. A new node v_i joins the communities to determine whether v_i is close to one community C_k, so the concept of membership degree of node and community is introduced, as follows:

$$S(v_i, C_k) = | E(v_i, C_k) | / d(v_i) \tag{2}$$

Among the formula: $|E(v_i,C_k)|$ represents the number of all connections between node v_i and community C_k, the $d(v_i)$ represents the degree of node v_i. Obviously, the greater $S(v_i,C_k)$ is, the closer connection between node v_i and community C_k is, and it describes the close degree between v_i and C_k. The initial small communities are formed by core nodes and its corresponding neighbor nodes.

C. Modularity Increment of Overlapping Communities

In actual network, some nodes belong to more than one community, so the research uses overlapping node. The overlapping node is not only belongs to a community but also belong to multiple communities, and is influenced by multiple communities.

The initial communities are divided by the above formula, and the algorithm proposed by Newman is applied to overlapping communities [17]. Assume the node u closely connects to the l communities, and is split into $u_1,u_2...u_l$ which are connected with l communities. The formula of modularity in overlapping communities is as follows:

$$Q = \frac{1}{2m} \sum_{k=1}^{l} \sum_{u_k, j=1}^{n} [r_{u_k j} a_{u_k j} - t_{u_k j} d(u_k) d(v_j) / 2m] \tag{3}$$

The $r_{u_k j}$ describes by membership degree from u_k and v_j to C_k. If u_k connects v_j, the $a_{u_k j}=1$, otherwise $a_{u_k j}=0$. $t_{u_k j}$ is a probability that u_k is connected to v_j and they simultaneously belong to C_k, and its formula is as follows:

$$t_{u_k j} = [\sum_{p=1}^{n} S(u_k, C_k) S(p, C_k) / n] \bullet [\sum_{p=1}^{n} S(v_j, C_k) S(p, C_k) / n] \tag{4}$$

In the formula: the $t_{u_k j}$ is a product between average probability of membership degree from u_k to C_k and average probability of membership degree from v_j to C_k. So the modularity of overlapping communities meets the above properties. The effect of identifying community is optimum when Q reaches the maximum. However, it is difficult to directly determine whether Q has reached the maximum. Therefore, the research uses modularity increment ΔQ, and its meaning is a range of increase or decrease between Q_n and Q_{n-1}, and its expression is as follows:

$$\Delta Q = Q_n - Q_{n-1} \tag{5}$$

When the $\Delta Q > 0$, the modularity increases, otherwise, the modularity decreases. If the ΔQ is zero, the modularity is maximum, so the algorithm is over.

III. Parallel Computing Framework Based on MapReduce

A majority of time is occupied by calculating node clustering coefficient and edge aggregation coefficient to seek core nodes and calculating membership degree between the split nodes and corresponding communities. Therefore, the research designs a parallel community identification algorithm based on Map/Reduce [18] to improve the computational efficiency.

A. Expression for MapReduce Parallel Computation

For overlapping communities with overlapping nodes, the first step is to find the core node. The input of key-value pairs is <*key, value*>, and the *key* is an offset at the beginning of text and the *value* is a 5-tuple <v_i, $c(i)$, $C(ij)$, *amt*, $s(v_i)$>, and the *amt* is metadata including important information about entity, and the $s(v_i)$ is a current status of node.

B. MapReduce Computing Framework

1) Map phase

a) The network graph is decomposed into key-value pairs, $G=(V,E)\rightarrow[(key,value)_1,(key,value)_2\ldots(key,value)_n]$;

b) Traverse the nodes in the graph, and calculate the degree of nodes d(vi);

c) The core node Center(i) is found by calculating point clustering coefficient and edge clustering coefficient;

d) Calculate relationship between core node and corresponding its neighbor nodes to form initial small communities;

e) Calculate membership degree between newly added node u and initial small community C to identify community from local community to global community;

2) Reduce phase

a) Traverse the formation of the community map to seek overlapping nodes;

b) The overlapping node is split into $u_1,u_2\ldots u_n$ *to be calculated membership degree* $S(u_k, C_k)$;

c) If new communities are formed by overlapping nodes, then multiple communities merge;

d) Calculate the new modularity increment ΔQ;

e) If $\Delta Q > 0$, *return a) step, identifying community is best until* $\Delta Q < 0$ *and the algorithm is over.*

IV. EXPERIMENT DATA

A. Citation Network Data Set

The experiment used Database Systems and Logic Programming (DBLP) for testing network community, and the data set had following characteristics in Table.1:

TABLE I. MAIN FEATURES OF DATA SET

Data set	Nodes(/10^3)	Edges(/10^3)	Storage(/GB)
DBLP	211.867	1296.522	0.253
DBLP V1	1.266	5.193	0.051
DBLP V2	8.123	43.965	0.124
DBLP V3	44.074	368.105	0.354

The DBLP was a subset of graph data in coauthored network, including title, author, references and some other information. In experiment, DBLP was transformed into a highly abstract network topology, where nodes represented papers and edges represented citation relations between papers. Based on the DBLP data set, a certain number of nodes and edges were selected to form DBLP V1. On the basis of DBLP V1 data set, the number of nodes and edges was increased to gradually form DBLP V2 and V3.

B. Experimental Result

First, the algorithm of overlapping community was tested on DBLP V1 data set and four distributed computing nodes. Then the scalability of parallel computing framework was verified by increasing data of DBLP V2 and V3. Finally, the running times illustrated that the parallel framework algorithm was better than serial algorithm and CONGA algorithm [19]. The Tp, Ts and Tc respectively represented running time of the parallel algorithm, serial algorithm and CONGA algorithm. As shown in Table.2, compared with other algorithms, the parallel algorithm saved running time with the increase in the amount of network data.

TABLE II. COMPARISON OF RUNNING TIME OF THREE ALGORITHMS

Data set	Tp(s)	Ts(s)	Tc(s)
DBLP V1	0.278	0.754	0.972
DBLP V2	1.535	5.964	7.335
DBLP V3	4.761	21.438	29.812

In addition to using the module increment to identify overlapping community networks, the experiment introduced overlapping degree, which was a measure of overlap parts between overlapping node and each community which it belonged to. The research took into account the close degree of split node u_k and overlapping communities, the following formula is shown:

$$O(u_k, C_k) = \frac{S(u_k, C_k)}{\sum_{i=1}^{l} S(u_i, C_i)} \tag{6}$$

In the formula: The denominator was a sum of membership degree which split node u_k belongs to each community, and the numerator was a membership degree which split node u_k belongs to the current community. Therefore, the greater $O(u_k,C_k)$ was, the higher overlap degree between overlapping node and the current community was.

According to the above formulas, the experiment calculated overlap degree and modularity increment to test relationships between them and iterations, and the results were shown in Figure 1.

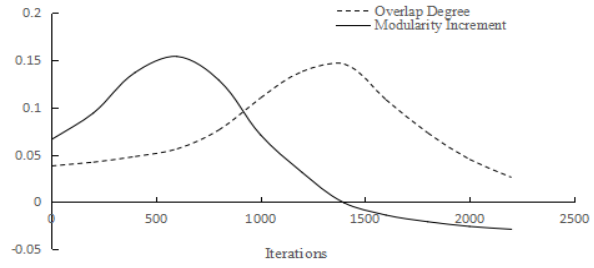

FIGURE I. RELATIONSHIP BETWEEN MODULARITY INCREMENT AND OVERLAP DEGREE AND ITERATIONS

As shown in Figure 1. According to the change of modularity and iterations, the modularity increment ΔQ was zero when iteration approximately was 1400th. After that, the modularity increment was negative. At the same time, the overlap degree $O(u_k,C_k)$ reached a peak value at about 1400th iterations with the relationship between overlap degree and iterations, and it decreased rapidly later, and it could be seen that the number of overlapping communities reduced clearly after the algorithm had reached a certain value. Therefore, the effect of actual complex network identification was optimal when the overlap degree reached a peak value and simultaneously the modularity increment is zero state about 1400th.

V. CONCLUSION

The research focused on the characteristics of complex overlapping community networks to use modularity increment and overlap degree to better identify community, and the algorithm saved time and improved efficiency through the use of parallel computing framework. The next step in the research is to solve the overlapping community in directed graph, because the relationship between node and node is directional, and contains important information, and it probably truly reflects the transitional information of nodes and close relationship. It can be more efficiently identify community with calculating direction in overlapping community.

ACKNOWLEDGMENT

This project was supported by the Funding Project for Natural Science Foundation of China (Grant No. 61671070), the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (Grant No.ICDD201608) and the Project of Computer Teaching Reform of Beijing Information Science and Technology University (Grant No. 5111623409).

REFERENCES

[1] Barabasi A, Albert R, Jeong H, Bianconi G. Power-law distribution of the World Wide Web. Science, 2000,287(5461):2115.

[2] Zhang Bo, Fu Yan. Research on community discovery algorithm in directed network [D]. Xi'an, University of Electronic Science and Technology, 2013, 6, 1-3.

[3] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12):7821-7826.

[4] Newman M. Modularity and community structure in networks[J]. PNAS, 2006,103(23):8577-8582.

[5] Yin Junsong, Wang Xiaoming, Zhang Xinqiang et al. The discovery algorithm for complex networks based on spectrum analysis [J]. radio communication technology, 2016,42 (5): 48-52.

[6] Papadakis H C, Panagiotakis C, Fragopoulou P. Locating communities on graphs with variations in community sizes[J]. Journal of Supercomputing,2013,65(2):543-561.

[7] Yu Hai, Zhao Yuli, Cui Kun, Zhu Zhiliang. A community algorithm based on cross-entropy discovery [J]. Journal of Computers .2015, 38 (8) 1574-1581.

[8] Liu Chao, Zhu Fuxi, Gan Lin. Based on label propagation probability overlapping community discovery algorithm [J]. Journal of Computers. 2016, 39(4): 717-729.

[9] Fortunato S.Community detection in graphs [J]. Physics Reports, 2010, 486 (3-5): 75-174.

[10] Cheng Xueqi, Shen Huawei. Community structure of complex networks [J]. complex systems and complexity science, 2011, 8 (1): 57-70.

[11] Mark E.j. Newman. The structure and function of complex networks [J]. SIAM Review, 2003,45: 167-256.

[12] Xie J, Kelley S, Szymanski B. Overlapping community detection in Networks: the state-of-the-art and comparative study [J]. ACM Computing Surveys, 2013,45(4).43:1-35.

[13] Zide Meng, Fabien Gandon, Catherine Faron-Zucker et al. Detecting topics and overlapping communities in question and answer sites [J]. Social Network Analysis and Mining, 2015,5(1):1-17.

[14] Xuewu Zhang, Huangbin You,William Zhu,et al. Overlapping community identification apprach in online social networks [J]. Physical A Statistical Mechanics and Its Applications, 2014,421:233 -248.

[15] Li Kongwen, Gu Qing, Zhang Yao, Chen Daoxu. A local community partitioning algorithm based on clustering coefficient [J]. computer science, 2010 (37): 46-53.

[16] Watts D, Strongatz S. The small world problem [J]. Collective Dynamics of Small-World Networks, 1998, 393: 440-442.

[17] Newman M. Modularity and community structure in networks[J]. Proceeding of the National Academy Science of USA,2006,103:8577-8582.

[18] Schuhmacher D, Vo B T, Vo B N. A consistent metric for performance evaluation of multi-object filters [J]. IEEE Transactions on Signal Processing, 2008,56(8):3447-3457.

[19] Gregory S. An algorithm to find overlapping community structure in networks [C]. Proc of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases(PKDD). Berlin: Springer-Verlag,2007:91-102