Network community partition based on intelligent clustering algorithm

Z.M. Cai [1]
[1]*Information Engineering College, Henan University of Animal Husbandry and Economy, Zhengzhou, Henan 450044, China*

Abstract

The division of network community is an important part of network research. Based on the clustering algorithm, this study analyzed the partition method of network community. Firstly, the classic Louvain clustering algorithm was introduced, and then it was improved based on the node similarity to get better partition results. Finally, experiments were carried out on the random network and the real network. The results showed that the improved clustering algorithm was faster than GN and KL algorithms, the community had larger modularity, and the purity was closer to 1. The experimental results show the effectiveness of the proposed method and make some contributions to the reliable community division.

Keywords: clustering algorithm, network community, node similarity, community division.

Citation: Cai ZM. Network community partition based on intelligent clustering algorithm. Computer Optics 2020; 44(6): 985-989. DOI: 10.18287/2412-6179-CO-724.

Introduction

With the development of society, various relationships in real life become more and more complex, forming various types of systems, such as information systems, transportation systems, power systems, satellite system [1], etc. If individuals in these systems are represented as nodes and relationships between individuals are represented as edges, a complex network can be obtained [2]. Complex network has a community structure, which contains the hidden relationship between nodes [3]. The division of community structure is one of the focuses of complex network research and has been widely concerned by researchers [4]. Golsifid et al. [5] designed a clustering model based on objective function, represented the degree of community membership with fuzzy numbers, and verified the effectiveness of the method through experiments. Bai et al. [6] proposed an iterative search algorithm, performed the initial partition before the starting of the algorithm to improve the quality of community division, and verified the efficiency of the method in community division through experiments on real networks. Zhang et al. [7] designed a fuzzy community division method to iteratively propagate the membership degree of all nodes and make full use of network topology information and found through the experiments that the method had low computational complexity and high performance. Zhang et al. [8] divided communities with spectrum clustering algorithm and established user similarity model which had a good performance in large-scale social network community division. This study mainly analyzed the application of intelligent clustering algorithm in community division, improved the Louvain clustering algorithm, and verified the reliability of the method through experimental analysis. The method is conducive to improving the efficiency and quality of community division and can be promoted and applied in practice.

Complex networks and communities

It is assumed that there is a network

$$G=\left(V,E\right), \tag{1}$$

where V stands for node and E stands for edge. The description of the network is as follows.

(1) Clustering coefficient: if node v has x neighbor nodes and E edges. The clustering coefficient is:

$$C_i=\frac{2E}{x\left(x-1\right)}, \tag{2}$$

and the average clustering coefficient is:

$$C=\frac{1}{N}\sum_{i}^{N}C_i\ . \tag{3}$$

Degree distribution: if adjacency matrix is

$$A=\left\lfloor a_{ij}\right\rfloor_{N\times N}, \tag{4}$$

then the degree is:

$$k_i=\sum_{j\in N}a_{ij}\ . \tag{5}$$

Degree distribution $p(k)$ represents the proportion of nodes with degree of k in the network,

$$p\left(k\right)=\frac{N\left(k\right)}{N}. \tag{6}$$

(2) Average path length: the average value of any distance between two nodes d_{ij},

$$d=\frac{1}{1/2N\left(N-1\right)}\sum_{i\geq j}d_{ij}\ . \tag{7}$$

The community structure in the network is shown in Fig. 1. Similar nodes exist in the same community, and

there are many connections between nodes in the same community. At present, there are many methods about community division, including graph segmentation [9], label propagation [10], hierarchical clustering [11], matrix spectrum [12], etc.

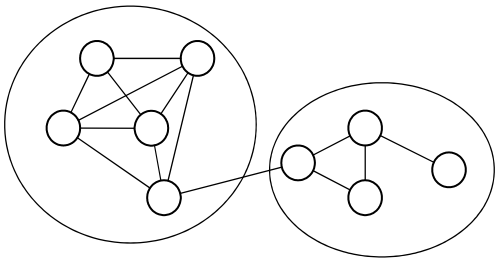

Fig. 1. Community structure in the network

When processing community division, the evaluation of the division result is an important part. Generally it is evaluated by modularity and purity.

(1) Modularity: for

$$G=(V,E), \tag{8}$$

the number of nodes is n, the number of edges is m, the degree is k, and the expectation of the edge between node i and j is $k_i k_j / 2m$, then the modularity is:

$$D=\frac{1}{2m}\sum_i\sum_j\left(A_{ij}-\frac{k_i k_j}{2m}\right), \tag{9}$$

and its value is between -1 and 1; the larger the value is, the better the division result is.

Purity: If the division result of community is:

$$H=\{H_1,H_2,\cdots,H_o\} \tag{10}$$

in the real situation, where o stands for the number of communities, then there is $H_i\cap H_j=\varphi$ for any $i\neq j$. The division result of the algorithm is:

$$G=\{G_1,G_2,\cdots,G_o\}. \tag{11}$$

The purity is:

$$P(G,H)=\frac{\sum_j \max_i |H_j\cap G_j|}{n}, \tag{12}$$

and its value is between 0 and 1; the closer to 1, the closer the results to reality.

Community partition based on intelligent clustering algorithm

1. Clustering algorithm

Louvain algorithm is a typical hierarchical clustering algorithm, which is simple and efficient, and has advantages in large-scale network processing [13]. The steps of the algorithm are as follows:

(1) The initialization nodes are independent communities, i.e., the number of nodes = the number of communities.

(2) Node i is added to the community of neighbor node, and increment ΔQ of modularity at that moment is calculated. The maximum value $\Delta Q_{\max}$ is taken. If

$$\Delta Q_{\max}>0, \tag{13}$$

the node will be added to the community, otherwise it will remain unchanged.

(3) Step (2) repeats until the modularity no longer changes.

(4) The community obtained in step (3) is taken as the supernode, and the above steps repeat until the modularity remains unchanged.

Lougain algorithm has a high running speed and a better division of small-scale communities. However, it only considers the link information between nodes, which makes the compactness of nodes decline and may affect the accuracy of the division results. Therefore, this paper improves the algorithm based on the similarity of nodes.

2. Node similarity

There are many methods to calculate the similarity of nodes, and the classic ones are some algorithms which consider the number of common neighbors as the standard:

(1) common neighbor (CN):

$$s(i,j)=|\mathrm{N}(i)\cap\mathrm{N}(j)|, \tag{14}$$

(2) cosine similarity (Salton):

$$s(i,j)=\frac{|\mathrm{N}(i)\cup\mathrm{N}(j)|}{\sqrt{k_i\times k_j}}, \tag{15}$$

(3) HPI (hub depressed index):

$$s(i,j)=\frac{|\mathrm{N}(i)\cap\mathrm{N}(j)|}{\min\{k_i\times k_j\}}, \tag{16}$$

where N(i) stands for a set of neighbor nodes, k refers to the degree, and $|\mathrm{N}(i)\cap\mathrm{N}(j)|$ refers to the number of common neighbors.

There are also algorithms based on the degree of common node, mainly Adam-Adar (AA) [17] and Resource Allocation (RA). According to AA algorithm, the larger the degree of node, the higher the contribution degree. The calculation method is as follows:

$$s(i,j)=\sum_{c\in\mathrm{N}(i)\cap\mathrm{N}(j)}\frac{1}{\lg k_c}, \tag{17}$$

where c stands for the common neighbor node and k_c refers to the degree of c.

According to RA algorithm, the common neighbor node can distribute its resources equally to neighbor nodes, then its similarity is the total resources received from i by node j; the calculation method is:

$$s(i,j)=\sum_{c\in\mathrm{N}(i)\cap\mathrm{N}(j)}\frac{1}{k_c}. \tag{18}$$

In this study, RA algorithm is selected to obtain node similarity. In order to further improve the accuracy of community division, RA algorithm was improved. Based on the relationship between common neighbor nodes, an improved RA algorithm is obtained:

$$s(i,j)=\frac{b_t}{b_T}\sum_c\frac{1}{k_c}, \tag{19}$$

Where

$$c\in\begin{cases}\mathrm{N}(i)\cap\mathrm{N}(j), \textit{the side of i is not directly connected with that of j}\\ \mathrm{N}(i)\cap\mathrm{N}(j)\cup\{i,j\}, \text{the side of i is directly connected with that of j}\end{cases}, \tag{20}$$

$$\frac{b_t}{b_T}=\frac{b_t}{T(T-1)/2}, \tag{21}$$

where b_t stands for the total number of edges in the network which is composed of node i and j and their common neighbor nodes, b_T stands for the maximum total number of edges expected, and T stands for the number of nodes.

3. Improved clustering algorithm

The steps of improving the clustering algorithm are as follows:

(1) the similarity between nodes was calculated;

(2) each node was divided as a community and added to the community of neighbour node; if the modularity value increased, that node was added to the community with the most increase;

(3) after the initial division, each community was taken as a node, and the sum of the similarity of the connected nodes between the two communities was taken as the edge; step (2) repeats until the modularity remained unchanged.

Experiment and analysis

1. Experimental data set

(1) Random network R (4, N, 16, 0.8) was used, where 4 refers to the number of communities, N refers to the number of nodes, 16 refers to the average degree of nodes, and 0.8 refers to the tightness of node connections.

(2) Real data sets [15] included YouTube (user and user relationship network), DBLP (author partnership network), Zachary (karate club membership network) and an American university football team network, as shown in Table 1.

Table 1. Real data set

Data set	Number of nodes	Number of edges
YouTube	1134890	2987624
DBLP	317080	1049866
Zachary	34	78
Rugby team	115	613

2. Experimental results

Firstly, the community partition method proposed in this study was tested on random network *R*and compared with the classical GN algorithm [16] and KL algorithm [17]. The network scale was adjusted through setting the value of N in R (4, N, 16, 0.8). The community partition time of different algorithms is shown in Fig. 2.

It was seen from Fig. 2 that GN had the longest computing time and the slowest computing speed among the three algorithms, KL algorithm was the second, and the algorithm proposed in this study had the least computing time and high efficiency.

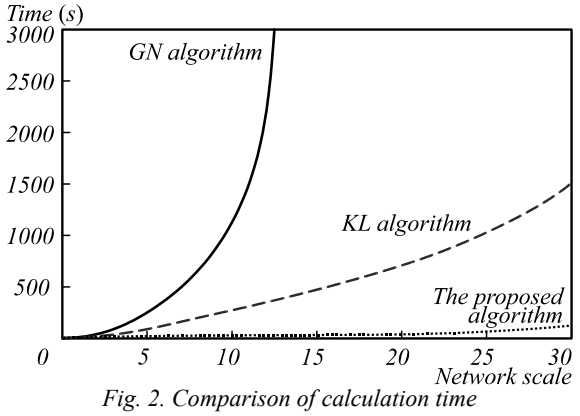

Fig. 2. Comparison of calculation time

The algorithm proposed in this study was applied to the real data set and compared with GN algorithm and KN algorithm. The comparison of modularity is shown in Fig. 3.

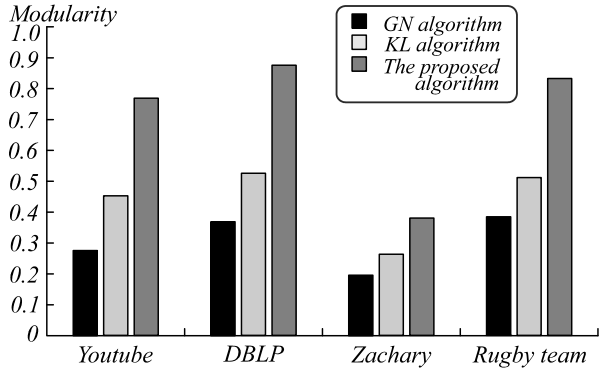

Fig. 3. Comparison of modularity

It was seen Fig. 3 that the modularity of the results divided by the algorithm proposed in this study was large. In the division of YouTube, the modularity of GM algorithm, KL algorithm and the algorithm proposed in this study was 0.276, 0.453 and 0.769 respectively. In the division of DBLP, the modularity of the three algorithms was 0.369, 0.526 and 0.876 respectively. In the division of Zachary, the modularity of the three algorithms was 0.196, 0.264 and 0.381 respectively. In the division of rugby team, the modularity of the three algorithms was 0.385, 0.512 and 0.833 respectively. It was found that the modularity of GN algorithm was the smallest, followed by KL algorithm and the algorithm proposed in this study, which showed that the algorithm proposed in this study had the best partition performance.

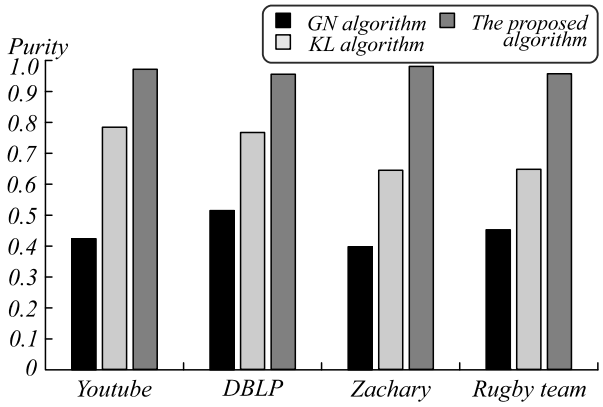

Fig. 4. Comparison of purity

Fig. 4 shows the results of purity comparison. In the division of YouTube, the purity of the three algorithms was 0.423, 0.784 and 0.972 respectively. In the division of DBLP, the purity of the three algorithms were 0.514, 0.767 and 0.956 respectively. In the division of Zachary, the purity of the three algorithms was 0.397, 0.645 and 0.981 respectively. In the division of rugby team, the purity of the three algorithms was 0.452, 0.648 and 0.957 respectively. It was found from the comparison that the purity of G N algorithm was the lowest, followed by K L algorithm and the algorithm proposed in this study, which showed that the partition result of the algorithm proposed in this study was closer to the actual situation.

Discussion

Community division is of great significance. For example, the online shopping network can recommend products according to hobbies of users to improve the marketing effect [18]; the social network can find hot spots and emotional tendencies of users, so as to control public opinion. Community division is conducive to a better understanding of network characteristics and effective mining of information in the network, which has great practical value [19].

There are many methods of community division. This study mainly analyzed the clustering method. Louvain algorithm has been widely used in community partition, and there are many researches on the improved Louvain algorithm. In this study, based on the method of node similarity, the Louvain algorithm was improved. First of all, it was found from the experimental results of random network that the algorithm proposed in this study had high computational efficiency, especially in large-scale networks. In the experiment of real network, modularity and purity were selected to evaluate the partition results. As shown in Fig. 3 and 4, the algorithm proposed in this study was superior to G N algorithm and K L algorithm in aspects of modularity and purity. The modularity value of the algorithm proposed in this study was larger, which showed that the result of community partition was better. The purity of the partition results of the algorithm proposed in this study was closer to 1, which showed that the result of community division of the algorithm proposed in this study was closer to the situation of real network and more in line with the reality.

The algorithm proposed in this study was an improvement of the clustering algorithm in community partition, but there are still some shortcomings, which need to be improved in the following aspects:

(1) the applicability of the method should be studied on overlapping networks;

(2) more calculation methods of node similarity should be studied to find out the most suitable algorithm for community division;

(3) the community division of dynamic network should be studied.

Conclusion

In this study, the network community was divided by the improved clustering method, and the experiment was carried out on the data set. The results showed that compared with G N and K L algorithm,

(1) the calculation time of the proposed algorithm was shorter on the random network;

(2) the modularity value of the proposed algorithm was larger and the partition effect was better in the real network;

(3) the purity of the proposed algorithm was closer to 1 on the real network, and the partition result was closer to the actual situation.

This study verifies the effectiveness of the improved clustering algorithm in the community division, and the method can be further promoted and applied in the network community division.

References

[1] Mostovoi JA. The two-phase operation in large-scale network. Computer Optics 2013; 37(1): 120-130.

[2] Pizzuti C. Evolutionary computation for community detection in networks: A review. IEEE Trans Evol Comput 2018; 22(3): 464-483.

[3] Reihanian A, Minaei B, Alizadeh H. Topic-oriented community detection of rating-based social networks. J King Saud Univ Sci – Computer and Information Sciences 2016; 28(3): 303-310.

[4] Rossetti G, Pappalardo L, Pedreschi D, Giannotti F. Tiles: an online algorithm for community discovery in dynamic social networks. Mach Learn 2016; 106(8): 1213-1241.

[5] Golsefid SMM, Zarandi MHF, Bastani S. Fuzzy duocentric community detection model in social networks. Int J Intell Syst 2015; 43(12): 177-189.

[6] Bai L, Cheng X, Liang J, Guo Y. Fast graph clustering with a new description model for community detection. Inf Sci 2017; 388-389: 37-47.

[7] Zhang H, Chen X, Li J, Zhou B. Fuzzy community detection via modularity guided membership-degree propagation. Pattern Recognit Lett 2016; 70: 66-72.

[8] Zhang H, Wu Y. Optimization and application of clustering algorithm in community discovery. Wirel Pers Commun 2018; 102(2): 1-12.

[9] Linares OAC, Botelho GM, Rodrigues FA, Neto JB. Segmentation of large images based on super-pixels and community detection in graphs. IET Image Proces 2016; 11(12): 1219-1228.

[10] Žalik KR. Community detection in networks using new update rules for label propagation. Computing 2016; 99(7): 1-22.
[11] Yin C, Zhu S, Chen H, Zhang B. A Method for community detection of complex networks based on hierarchical clustering. Int J Distrib Sens Netw 2015; 2015: 137.
[12] Chen PY, Zhang B, Hasan MA. Incremental eigenpair computation for Graph Laplacian Matrices: Theory and applications. Soc Netw Anal Min 2017; 8(1): 4.
[13] Cordeiro M, Sarmento R P, Gama J. Dynamic community detection in evolving networks using locality modularity optimization. Soc Netw Anal Min 2016; 6(1): 15.
[14] Mohan A, Venkatesan R, Pramod KV. A scalable method for link prediction in large real world networks. J Parallel Distrib Comput 2017; 109: 89-101.
[15] Konect. Source: ⟨https://west.uni-koblenz.de/konect⟩.
[16] Du W, He X. A common strategy to improve community detection performance based on the nodes' property. Commun Comput Inf Sci 2016: 355-361.
[17] Sun Y, Danila B, Josic K, Bassler KE. Improved community structure detection using a modified fine-tuning strategy. EPL 2009; 86(2): 28004.
[18] Yu H, Blair RH. A framework for attribute-based community detection with applications to integrated functional genomics. Pac Symp Biocomput 2016; 21: 69-80.
[19] Zhang X, Liu BQ, Wang XL. Research on community detection methods in complex network. CEA 2015; 13(2): 368.

Author's information

Zhongmin Cai (b. 1976) graduated from University of Electronic Science and Technology of China in 2013 and has gained the master's degree. He is working in Henan University of Animal Husbandry and Economy as an associate professor. He is interested in computer network. E-mail: *zhongmczm@163.com* .

Received March 25, 2020. The final version – May 8, 2020.