Extensible Spectral Clustering Method For Detecting Overlapping Communities in Large-scale Networks

Hong Zhong, Deyong Jiang

Yancheng Institute of Technology, Yancheng, 224000, China

Keywords: Network overlapping community detection; spectral clustering

Abstract: The networked description of complex systems in the real world enables people to more clearly recognize and understand the functions of the system and the future development of the system, and play a guiding role in the further improvement and optimization of the system structure, improving the production efficiency of the system, and saving costs. Community structure analysis is to decompose the overall structure of the network into several communities, so that the links between nodes in the community are dense and the links between nodes in the community are sparse. Traditional clustering methods are limited by the existing computing and storage conditions, which are often time-consuming and highly dependent on storage space. Therefore, the study of overlapping community auto-detection in large data of complex network is of great significance for more scientific and rational planning of complex large data network, optimizing network structure and ensuring network service quality. Finally, this is fundamentally different from traditional data mining techniques. It is characterized by large amount of data, numerous types, wide source channels, low value density, uneven data quality, fast growth, high aging, and large variability. In this paper, a spectral clustering integration algorithm for large-scale network overlapping community discovery is proposed, because the computationally expensive spectral clustering cannot meet the needs of large-scale network community discovery.

1. Introduction

The network community structure can describe the related characteristics of structure and function among the internal components of complex systems. It is of great theoretical significance and application value to discover the knowledge of community structure in various complex systems [1]. The networked description of complex systems in the real world enables people to recognize and understand the functions and future development of the system more clearly, and plays a guiding role in further improving and optimizing the system structure, improving the production efficiency of the system and saving costs. Complex networks have the characteristics of power law distribution, small average shortest path length, high aggregation coefficient and hierarchical community structure. Community structure analysis is to break down the overall structure of the network into several communities, making the links between nodes in the community dense and the links between nodes in the community sparse [2].Community structure is usually characterized by the close connection of nodes within the community and the loose connection of nodes between the communities. Traditional clustering methods, limited by the existing computing storage conditions, tend to be time-consuming and highly dependent on storage space [3].Therefore, how to carry out efficient clustering calculation under large-scale data is increasingly concerned by scholars. Therefore, the study of overlapping community auto-detection in large data of complex network is of great significance for more scientific and rational planning of complex large data network, optimizing network structure and ensuring network service quality [4].

Many natural or social complex systems, such as World Wide Web, social network and biological network, can be described by graph structure. By studying the community structure formed by interpersonal relationships, hobbies or other interactive processes in social networks, sociologists can reveal various hidden relationships between individuals and society in social networks. Finally, this point is essentially different from the traditional data mining technology [5]. Large amount of data, various types, wide sources, low value density, unbalanced data quality, fast

  DOI: 10.25236/mmmce.2021.048

growth, high timeliness and great variability. As the basis of studying the network structure, revealing the community structure in the network is very important for studying the function of the network and analyzing the composition and structure of the network [6]. Detect overlapping communities in large data from complex networks based on computational results [7].However, this method does not divide the community structure in detail, resulting in low accuracy of detection results .In recent years, complex networks have attracted much attention in different fields such as physics, sociology and computer science. How to discover valuable potential relationship patterns and community structure from these complex information networks is of great significance to Internet marketing, information recommendation, information dissemination control, crime network detection, computer virus dissemination and other aspects.

2. Spectral clustering ensemble algorithm

2.1. Generation of individual clusters

Choosing KASP algorithm to generate individual clustering is mainly due to two considerations, but using Nystrom rank reduction strategy can not explicitly give the direct correlation between data reduction and clustering accuracy. Because the algorithm requires users to set parameters such as upper bound of community number, threshold of relative distance and approximate weight under community according to specific data set, it is a delicate problem for users. Therefore, how to make the algorithm adaptively adjust these parameters in the iterative process is the focus of our future research [8]. They pointed out that the two basic mechanisms of scale-free networks are: growth and preferential connection. Growth and preferred connections indicate that the network has temporal characteristics. The network is in a dynamic process, and new joined nodes tend to form connection relationships with larger degree nodes. The algorithm first obtains the cluster centers of each class using k-means and uses the class centers as anchors. These anchors are clustered using a spectral clustering algorithm to get the clustering result [9].By defining the fitness function of nodes, the fitness of all neighbor nodes of the community to the community is calculated, and the node with the greatest fitness to the community is selected to join the community in order to achieve community division [10].This algorithm allows nodes to be grouped into multiple communities, thus enabling the discovery of overlapping communities.

2.2. Selection of individual clusters

There are two defects in the traditional framework of cluster integration learning: 1. Lack of individual cluster selection, all individual clusters are directly regarded as member clusters, and integrated by equal treatment. 2. Discover overlapping communities by optimizing the modularity of overlapping community structure. Carrot2 provides API interfaces for many text clustering algorithms such as Lingo, STC and so on. The module completes the basic process of hot spot discovery. As shown in Figure 1.

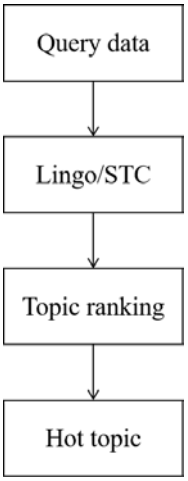

Figure 1 Hot topic discovery process

In the process of studying complex networks, degree distribution is usually used as an important parameter to measure the structural characteristics and evolution process of networks. After the anchor points are obtained, the data samples are transformed into anchor points by the base coordinate transformation operation. By calculating the similarity between community pairs, select the community with the greatest similarity to merge, and recalculate the similarity between these new communities and other communities until there is only one community in the network. Cluster integration is the application of the idea of integrated learning in the field of unsupervised learning. Its core is to merge different clusters of the same dataset into one result cluster which depicts the true distribution of data better than a single cluster. Moreover, the relationship between clustering accuracy and individual clustering diversity is not simple and positive, and it needs to be determined by the distribution characteristics of the data samples.

3. The discovery of network overlapping communities

3.1. Modularization of overlapping communities

Social network, as a typical complex network, has long attracted the attention of sociologists. Social network usually refers to the collection of social actors and their relationships. Then, we will introduce an improved locally constrained linear regression method based on anchor point representation. When a wanderer enters a community, the probability of choosing nodes in the community is much higher than that of choosing nodes outside the community. The higher the ratio of overlapping nodes, the higher the degree of overlap between communities, which means that the boundary between network communities is blurred. Based on this interactive process of Wed2.0, a network model between user topics can be built. As shown in Figure 2.

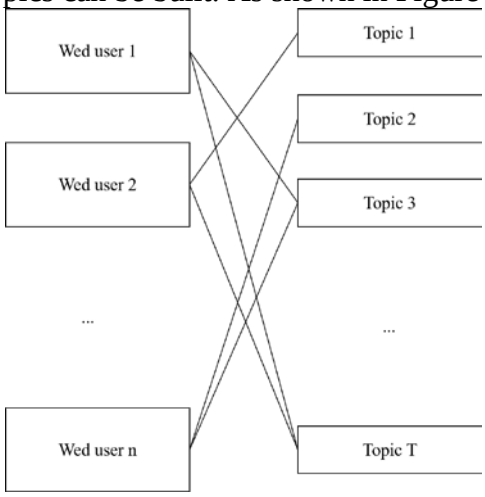

Figure 2 User-topic network model

Whether it is low modularity or high fuzziness, it will increase the difficulty of community mining, which will reduce the effectiveness of community mining algorithm results. With the increasing proportion of overlapping nodes in the simulation network, the probability of overlapping communities increases, and the detection difficulty also increases. This tag update strategy usually results in a large-scale community structure due to the randomness of the selection, even when the nodes of the entire network are assigned to the same community. Compared with the existing spectral clustering algorithm, the proposed spectral clustering algorithm requires fewer parameters, only two parameters need to be adjusted - the number of anchors and the number of anchor neighbors.

3.2. Rough clustering

Spectral clustering for unsupervised learning When clustering data, the most commonly used method is to cluster feature vectors embedded in space by means of mean algorithm, although the mean algorithm is simple and intuitive and has low computational complexity. However, it can only

divide data hard, and cannot meet the needs of discovering overlapping communities. In a coarse-grained model, the population is divided into subpopulations, each of which performs a simple genetic algorithm separately. The algorithm uses a coarser-grained evolutionary model of multiple populations, and the communication topological relationships among populations use a ring structure. As shown in Figure 3.

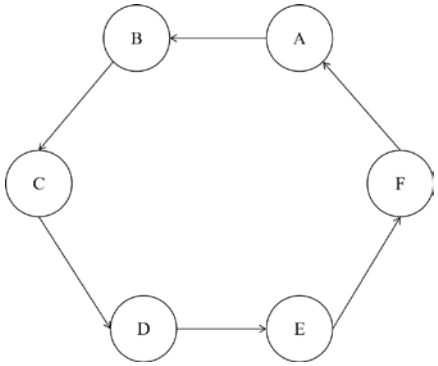

Figure 3 Network topology diagram of population migration

The local similarity representation uses the prior knowledge embedded in the given similarity function, that is, people's expectation for the intrinsic connection of data, and this description can sometimes make the data sparse. On the contrary, if the paths of the random walk process from these two nodes are very different, then this edge is probably an edge between communities. Since it is difficult to obtain prior knowledge about scale parameter selection from real data, which greatly limits the practical application of spectral clustering, is SCEA, as an integrated algorithm of spectral clustering, facing the same problem? The detection accuracy of the proposed method is the best in both small and large community complex networks, and the detection results are always stable. The other two methods are heavily affected by the proportion of overlapping nodes, with high and low fluctuations.

4. Conclusions

Theoretical and simulation tests show that the automatic detection method of overlapping communities based on fuzzy spectral clustering has high detection accuracy and efficiency. Based on the theory of complex network, this paper introduces and analyzes several representative community discovery algorithms and evaluation methods of network community structure. In the face of large-scale data sets, the traditional clustering method is limited by the existing computing storage conditions, which is often time-consuming and highly dependent on storage space. Because a node may have multiple edges associated with it, it is divided into different communities with different edges. Nodes are also divided into different communities so that overlapping nodes belonging to multiple communities can be found. Community pattern mining is an important topic in the study of complex networks, which is essentially a cluster analysis problem of network nodes. Common community discovery algorithms based on individual spectral clustering have some drawbacks, such as high time complexity and sensitivity to scale parameters when constructing similarity matrices, which make it impossible to effectively discover the real community structure implied in the network. At the same time, it has good stability performance, and has a guiding role for more scientific and rational planning of complex large data networks, optimizing network structure, and guaranteeing network service quality.

References

[1] Cui Yixin, Chen Xiaodong. Large-scale spectral clustering parallel algorithm optimized by Spark framework. Computer Applications, vol. 40, no. 1, pp. 168-172, 2020.

[2] Yan Xiaopeng, Sun Yongbo. Fuzzy spectral clustering overlapping community discovery algorithm. Automation Instrumentation, vol. 37, no. 3, pp. 27-29, 2016.

[3] Zhao Fei, Yu Benguo, Ji Qingbin. Research on spectral clustering community division method based on edge clustering coefficient. Journal of Central China Normal University (Natural Science Edition), vol. 189, no. 01, pp. 23-28, 2020 .

[4] Jiang Lifang, Su Yidan, Qin Hua. Parallel spectral clustering community mining algorithm based on CUDA. Shanxi Electronic Technology, vol. 000, no. 002, pp. 46-49, 2016.

[5] Zhang Xiaoqin, An Xiaodan, Cao Fuyuan. A binary network community discovery algorithm based on spectral clustering. Computer Science, vol. 46, no. 04, pp. 216-221, 2019.

[6] Zhang Haitao, Song Tuo, Zhou Honglei, et al. Research on knowledge aggregation method of virtual health community based on spectral clustering. Library and Information Service, no. 8, pp. 134-140, 2020.

[7] Wu Xiaolin, Fan Beibei. Research on complex product module discovery method based on improved spectral clustering algorithm. Metrology and Testing Technology, vol. 323, no. 04, pp. 72-75, 2019.

[8] Cui Yutong, Niu Qiang, Wang Zhixiao. Semi-supervised spectral clustering community discovery algorithm based on signal transmission. Computer Engineering and Design, vol. 39, no. 5, pp. 1201-1205, 2018.

[9] Zhang Wei, Qi Dehao, Chen Yunfang. Discovery of overlapping communities based on LDA model in large-scale networks. Journal of Nanjing University of Posts and Telecommunications: Natural Science Edition, vol. 8, no. 03, pp. 54-64, 2018.

[10] Teng Fei, Dai Rongjie, Ren Xiaochun. Parallel Algorithm for Discovery of Overlapping Communities in Complex Networks. Journal of Southwest Jiaotong University, vol. 54, no. 001, pp. 211-218, 2019.