ISSN: 2688-9323

Community Discovery based on Network Representation Learning

Shaomei Kong

School of Economics and Management, Xidian University, Xi'an 710126, China

Abstract

[Purpose/Significance] On the basis of major community discovery methods and the development of machine learning today, this paper proposes a method to discover communities. [Method/Process] Firstly, each node in the network is represented by a vector through network representation learning. Secondly, by computing the similarity of node vectors, clustering algorithm is used to cluster node vectors, and the complex network is divided into communities. Finally, dimensionality reduction was performed by principal component analysis and the results of community partition were visualized. [Result/Conclusion] Through the verification of three classical data sets, this method can make full use of the node and edge information of the complex network, and represent the network as a low-dimensional dense vector, so as to cluster and discover the community.

Keywords

network representation learning, node vector, node similarity, community discovery.

1. Introduction

Complex network is an abstract and descriptive way of complex system. The nodes in the network are represented as the components of complex system, and the edges in the network represent the relationship between the components. Scholars found that the complex network shows a certain community structure, which reflects the similarity of individuals within the network. Studying the community of the complex network can more effectively analyze the overall characteristics of the network, so many scholars put into the community discovery exploration, and also put forward a lot of community discovery algorithms.

Community discovery can be understood as the clustering problem of complex networks. As long as the similarity between nodes can be calculated, complex networks can be clustered. How to better build the similarity between nodes is our research direction. If we can express the network topology space as vector space without losing the structure information and attributes of the network, we can cluster by calculating the distance between nodes. In the past, the vector representation of complex networks usually used adjacency matrix. If two nodes are connected, the corresponding matrix element is 1, otherwise it is 0. However, the adjacency matrix is a high-dimensional sparse vector. The high-dimensional sparse vector can not get satisfactory results in the traditional clustering method, and consumes a lot of running time and computing space, so the adjacency matrix is not the best method to calculate the similarity between network nodes. With the development of machine learning and natural language, network representation learning algorithm can use random walk path to treat nodes in the network as "words" and walk path as "context" of words, so that the information of nodes and edges in the network can be represented as low-dimensional dense vectors after training, and nodes can be mapped into vector space to calculate the phase between nodes Similarity clustering is used to find the communities of complex networks.

ISSN: 2688-9323

2. Relevant Research

Compared with the classical community discovery algorithms, there are GN algorithm [1], Newman fast discovery algorithm [2], label propagation algorithm [3], and spectrum analysis algorithm. These algorithms have corresponding advantages and disadvantages.GN algorithm has high complexity and is not flexible enough; Newman quickly finds that the algorithm has significant efficiency but data redundancy; the algorithm of label propagation algorithm has fast convergence speed, but the result is unstable and the accuracy is relatively low. Spectral analysis method was first used to solve the problem of Graph Segmentation, which has high time complexity and can only be divided into two communities at a time. Based on these algorithms, many community discovery algorithms have been proposed by scholars at home and abroad.

When calculating the similarity of network nodes, there are metrics based on network structure information and metrics based on random walk of network. In recent years, there are many improved community discovery algorithms based on node similarity, for example, Jiang Yawen [4] and others proposed a community detection algorithm based on node similarity; Wu Zhonggang [5] and others proposed a community discovery algorithm based on local similarity; Zhan Wenwei [6] and others proposed a hierarchical aggregation community discovery algorithm based on similarity module degree; Zhang Hu [7] and others proposed a community discovery algorithm based on multi-layer node similarity; Liu Miaomiao [8] and others proposed a weighted network community discovery method based on the similarity of common neighborhood nodes. However, these algorithms usually use the measure index based on the network structure information, that is, some common neighbors between nodes and nodes, which can not well represent the structure information of the network to calculate the similarity between nodes. In this paper, we use the node vector which is based on the random walk of the network and the learning representation of the network.

3. Community Discovery based on Network Representation Learning

3.1. Research Methods and Steps

(1) Construct node walk path

In this paper, we use the method of constructing node walk path in deepwalk and node2vec respectively. Deepwalk randomly samples a node VI as the root node and samples around until the maximum path length t is reached.Node2vec partial random walk method is different from deepwalk's random walk method in that it uses breadth first search and depth first search of graphs.Node2vec defines a two order random walk that guides the walk. It has two parameters p and Q, parameter p controls the possibility of revisiting the node immediately during the walk, and parameter q controls the breadth first search and depth first search of the random walk. By controlling the parameters of P and Q, the probability of one node transferring to other nodes is controlled.

(2) Construction of node vector

After getting the walk path of nodes, the nodes in the network are regarded as "words" and the walk path of nodes is regarded as the context of words. The skip gram model in word2vec can be used to learn the representation of network nodes. The skip gram model includes input layer, hidden layer and output layer. When the input is text, the input word w (T) is given to predict the context of the word. For the network, the node n (T) is given to predict the edge of the node. The structure is shown in Figure 1.For text, the word vectors from semantically similar words are very close in vector space, and for network, the node vectors from neighbor similar nodes are also very close in vector space.

ISSN: 2688-9323

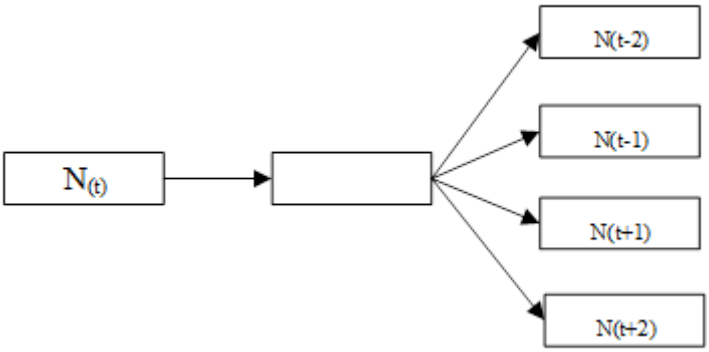


Figure 1: skip gram model

(3) Node vector clustering

After getting the 128 dimensional node vector space of the network, the 128 dimensional node vector of the network is shown in Figure 2. We can use clustering algorithm to discover the community according to the node similarity of the network. In the previous research, some scholars found that for the complex network community discovery problem, if we can choose a good method to construct the network node similarity, Then any clustering algorithm based on similarity matrix can quickly and effectively partition network communities [4].The clustering method used in this paper is kmeans clustering method which uses distance as the similarity evaluation index. The closer the distance between two objects, the greater the similarity and the higher the possibility of belonging to the same community.

```
8 -0.10097184 -0.0009648778 -0.01786216 -0.042192306 -0.15038243 -0.05555245 0
7 -0.08768378 -0.030035611 0.038054224 -0.14419249 -0.26386607 -0.14109139 0.0
6 -0.13897444 0.0061568446 0.017565858 -0.093612276 -0.32781386 -0.05531803 -0
5 -0.14650376 0.014285455 -0.0054417355 -0.07987969 -0.29574227 -0.03221401 -0
4 -0.09115475 -0.00088649953 -0.009296978 -0.07395297 -0.21970096 -0.013099957
3 -0.07521432 -0.04390147 0.024448046 -0.14130022 -0.21090578 -0.1028244 0.016
2 -0.09351292 -0.0088636335 0.020767227 -0.10549673 -0.20785584 -0.08621167 0.
1 -0.060624514 -0.028542371 0.044807237 -0.10387218 -0.18953402 -0.12081167 -0
```

Figure 2: 128 dimensional vector of nodes

(4) Dimensionality reduction and visualization of node vectors

Since the node vector obtained by network representation learning is 128 dimensions, the results of community discovery can not be visualized. Therefore, the principal component analysis method is used to reduce the dimension of 128 dimensional node space vector data, and then the community division results of the above steps are visualized.The 128 dimensional node space vector is reduced to 2-dimensional plane space vector, which is convenient for us to visualize the node vector and display the results of community division.

4. Experimental Results

In the experiment, the accuracy of the final results compared with the real results was used as the evaluation index. We selected three classic real datasets: Zachary karate club network, books on politics, and dolphin network. The three real networks have clear community structure, which is convenient for us to compare with the experimental results.

4.1. Karate Club Network

Karate club network is a classic data set in the field of social network analysis, which is composed of 34 nodes and 78 sides, as shown in Figure 3.Each node in the network represents a member of the club, while indicating that members of the club have frequent friends. The real result of the network is two communities. After network representation

ISSN: 2688-9323

learning is used to represent the network as node vector, kmeans is used to aggregate the network into two categories, and the optimal clustering accuracy is 100%.

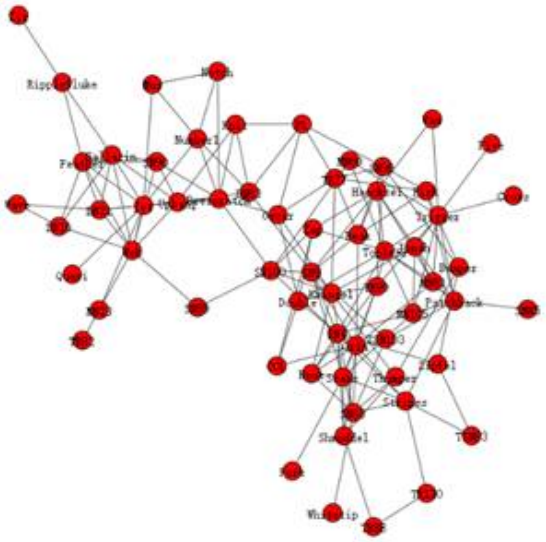


Figure 3: karate club network visualization

The 128 dimensional node space vector of karate club network is reduced to 2-dimensional plane space vector, and the visualization result of community division is shown in Figure 4.It can be seen from the figure that the effect of community division is good, and it can be clearly divided into two types in the plane space, with high accuracy.

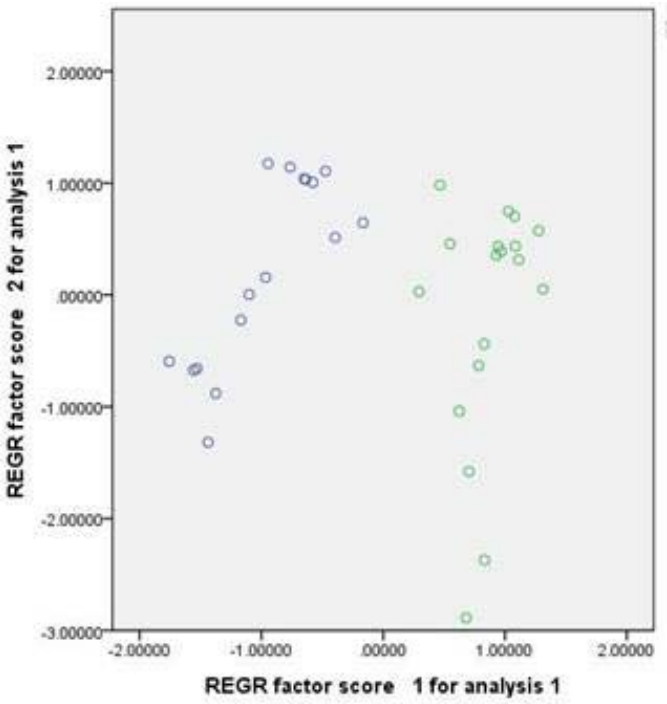


Figure 4: vector visualization of karate nodes in karate club network

4.2. Books on Politics

Books on politics is composed of 105 nodes and 441 sides, as shown in Figure 5.Each node in the network represents every book about American political theory sold by Amazon. Com, indicating that customers have bought both books at the same time. The real result of the network is three communities. After the network is represented as node vector by network

ISSN: 2688-9323

representation learning, kmeans is used to cluster into three categories. The optimal clustering accuracy is 84.76%.

Figure 5 :visualization of books on politics

The 128 dimensional node space vector of books on politics is reduced to 2-dimensional plane space vector, and the visualization result of community division is shown in Figure 6.It can be seen from the figure that the effect of community division is good, and it can be clearly divided into three categories in the plane space, with high accuracy.

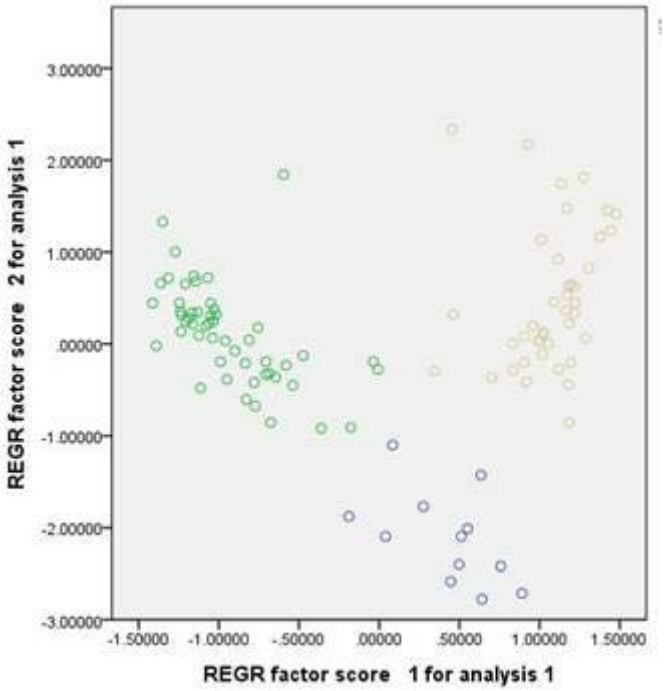


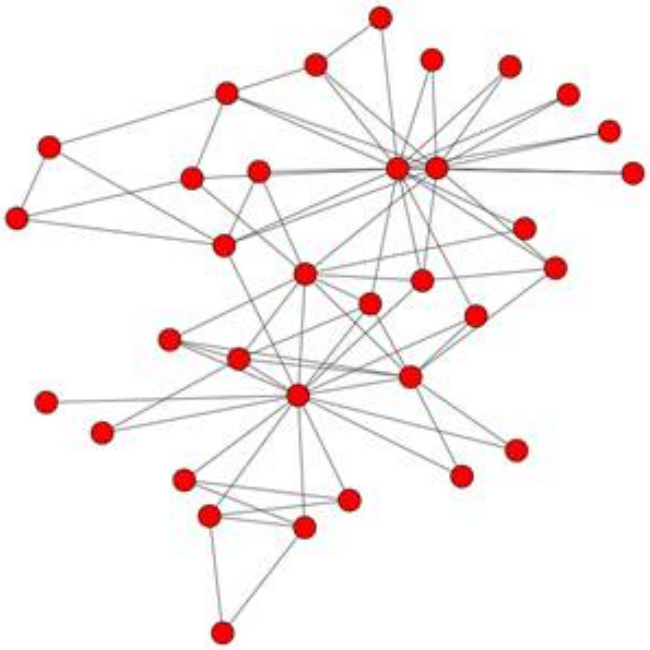
Figure 6: visualization of node vector of books on politics

4.3. Dolphin Network

Dolphin network consists of 62 nodes and 159 sides, as shown in Figure 7.Each node in the network represents this dolphin, while it represents frequent contact between dolphins.The real result of the network is two communities. After the network is represented as node

ISSN: 2688-9323

vector by network representation learning, kmeans is used to cluster into two categories. The optimal clustering accuracy is 98.387%.



;

Figure 7: dolphin visualization

The 128 dimensional node space vector of dolphin network is reduced to 2-dimensional plane space vector, and the visualization result of community division is shown in Figure 8. It can be seen from the figure that the effect of community division is good. It can be clearly divided into two categories in the plane space, and the accuracy is high.

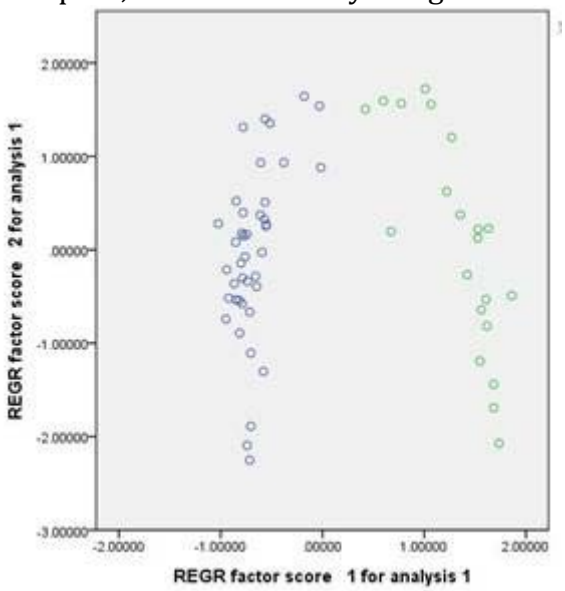


Figure 8: dolphin network nodes vector visualization

4.4. Comparison of Experimental Results

In this paper, the deepwalk, node2vec method is used to construct the node vector of the network, and the kmeans clustering algorithm is used to calculate the similarity of the node vector to cluster the node vector. The three classic networks are divided into communities. This method is compared with the accuracy of the classic community discovery algorithm GN algorithm and FN algorithm. The results are shown in Table 1.It can be seen from table 1 that the accuracy of clustering after using two kinds of network representation learning

ISSN: 2688-9323

algorithms is higher, among which the accuracy of node2vec clustering in karate club network karate and books on politics network of American political theory works is higher than that after deepwalk clustering, and the accuracy of other communities' discovery algorithms is quite effective.

Table 1. Comparison of accuracy of community discovery algorithm

	Karate club network	Books on politics	Dolphin network Dolphins
GN	0.9706	0.8381	1
FN	0.7353	0.8095	0.6935
Deepwalk + clustering	0.9706	0.8381	0.98387
Node2vec + clustering	1	0.8476	0.98387

5. Concluding Remarks

The development of machine learning provides us with a new way to solve the traditional problems. In the past, adjacency matrix was often used to represent the network, but the adjacency matrix is high-dimensional sparse and has some limitations. Network representation learning can use random walk to get the sequence of nodes to learn the vector representation of each node, so as to calculate the similarity of each node for clustering, which is impossible for traditional community discovery algorithm. The experimental results show that the accuracy of this method is high in the community. It can be seen that this method is suitable for most networks. Because of the low time complexity of this algorithm, it can also be used to study large-scale complex networks. By using network representation learning, network information can be transformed into low-dimensional and dense vectors, so that machine learning can be used to get more applications, such as node classification, link prediction, personalized recommendation and so on. In this paper, three classic datasets are used to verify the experimental results. The results show that the accuracy of the divided communities is high. The community discovery method in this paper is feasible and efficient. The network representation learning algorithm can express the topology of the network as a low-dimensional dense vector, and then cluster the communities by calculating the similarity between nodes.

References

[1] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.

[2] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.

[3] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(2):036106.

[4] Jiang Yawen, Jia Caiyan, Yu Jian. Research on network community detection algorithm based on node similarity [J]. Computer science, 2011,38 (7): 185-189.

[5] Wu Zhonggang, Lu Zhao. A community discovery algorithm based on local similarity [J]. Computer Engineering, 2016,42 (12): 196-203.

[6] Zhan Wenwei, Xi Jingke, Wang Zhixiao. Hierarchical aggregation community discovery algorithm based on similarity modularity [J]. Journal of system simulation, 2017,29 (05): 1028-1032 + 1040.

[7] Zhang Hu, Wu Yongke, Yang Zhizhuo, et al. Community discovery method based on multi-layer node similarity [J]. Computer science, 2018, 45 (1): 216-222.

[8] Liu Miaomiao, Guo Jingfeng, Ma Xiaoyang, et al. Weighted network community discovery method based on similarity of common neighborhood nodes [J]. Journal of Sichuan University (NATURAL SCIENCE EDITION), 2018, 55 (1): 89-98.