International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.209-218
http://dx.doi.org/10.14257/ijdta.2016.9.6.21

Hierarchical Community Detection Algorithm Based on Node Similarity

Jingke Xi, Wenwei Zhan and Zhixiao Wang*

College of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
zhxwang@cumt.edu.cn

Abstract

Louvain algorithm is a community detection algorithm based on modularity optimization. It is extremely fast, but the accuracy of detecting communities needs to be improved. This is because modularity of Louvain only considers link information between nodes and neglects the effect of the surrounding neighbor nodes, leading to decreased tightness between nodes in the same community and consequently affects accuracy. To solve this problem, by introducing node similarity to improve modularity function of Louvain algorithm, we propose a hierarchical community detection algorithm based on similarity (SHC).We adopt the Normalized Mutual Information to evaluate the accuracy of the algorithm and conduct experiments on the real network and the LFR synthetic network. The results show that the improved algorithm is more accurate, compared with Louvain and Newman Fast Algorithm.

Keywords*: Louvain algorithm; SHC algorithm; Modularity; Similarity; Community detection*

1. Introduction

Nowadays, there are many networks such as electronic commerce networks, copyright cooperation networks and online social networks in the real society. These networks are described as complex networks for their complex internal structure. In a complex network, a node represents a member and an edge represents relationship between two members. Community structure is an important property of complex networks. Research shows that [1]: nodes are densely connected in the same community and sparsely connected in different communities. Community structure in complex networks has important theoretical and practical value. Detecting community structure is conducive to understanding and using networks. So people proposed a lot of community detection algorithms.

Community detection algorithms are usually divided into three categories: methods based on graph partition, hierarchical clustering methods and extremal optimization methods. One representation of traditional graph partition methods is Kernighan-Lin algorithm[2].It first defines a gain function Q which represents difference value between sum of edges within communities and sum of edges between different communities, then divides all nodes into two communities with the same size, and then continuously exchanges nodes between two different communities in order to optimize the value of Q until all nodes in either of two communities are exchanged, finally iterates over sub-communities until getting the given number of communities. Hierarchical clustering methods[3] have two types that are agglomerative methods and divisive methods. Agglomerative methods regard each node as a separate community at first, then continuously merge communities according to the given rules and finally get the result of community detection. Divisive methods are contrary to that mentioned above, they regard all nodes as one community at first, then continuously delete edges in the community

ISSN: 2005-4270 IJDTA

according to the given rules and finally get the result of community detection. One typical method of divisive methods is Girvan-Newman algorithm [4].It uses edge betweenness to measure the importance of edges and continuously deletes the edges whose betweenness are the biggest until there is no edge to be deleted. An extremal optimization method defines an objective function at first, then looks for optimal division of community structure through continuously optimizing objective function. There is a typical algorithm called Fast-Newman algorithm [5], it adopts Modularity as objective function and gets optimal outcome when the value of Q is the biggest.

Louvain algorithm [6] is a hierarchical method that optimizes modularity [7]. It is easy to implement due to its intuitive steps. Besides it runs very fast so that it can deal with large-scale networks. What's more, it can avoid the so-called resolution limit of modularity [8] in some degree thanks to its hierarchical nature. The algorithm is recommended as the best performance community detection algorithm based on modularity optimization by well-known scholar Fortunato [9]. However, the modularity of Louvain algorithm only considers the link information between nodes and ignores the surrounding neighbor nodes, which leads to the tightness of nodes in the same community decreasing. As a result, it affects the accuracy of the final result. This paper introduces node similarity and makes appropriate improvements, and then redefines the modularity, finally we propose a similarity-based hierarchical clustering algorithm.

2. Similarity-Based Hierarchical Clustering Algorithm

2.1. Node Similarity

Node similarity is usually described as follows: in a network, if two nodes have similar neighbors, then they are similar. Methods of calculating node similarity are usually divided into three categories that are attribute-based methods, methods based on global link and methods based on local link. An attribute-based method [10] constructs attribute vectors for every node, then maps them to multi-dimensional space, and then calculates similarity through euclidean distance or other approaches. A method based on node link calculates similarity through link between nodes such as shortest path length or number of independent paths between nodes. The difference between methods based on global link and local link is that the former calculates similarity of all nodes while the latter only calculates similarity between nodes that are directly connected. As a result, the former is more accurate but has much larger computing. People usually choose the latter in practice. There are some examples of methods based on local link such as cosine similarity [11], Jaccard similarity coefficient, Dice similarity and so on. This paper adopts cosine similarity, and it can be calculated by the formula as follows:

$$\sigma(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)||\Gamma(v)|}} \tag{1}$$

The above function calculates similarity between two nodes through their common neighbors, the more common neighbors they have, the greater their similarity is. But there is a problem that when two directly connected nodes have no common neighbors, their similarity is 0 while it is obviously not true. Reference [12] takes two nodes which are directly connected into account when considering their common neighbors, its similarity is defined as follows:

$$S(i, j) = \frac{\sum_{V_e \in St(i) \cap St(j)} \frac{1}{k_e}}{\sqrt{\sum_{V_e \in St(i)} \frac{1}{k_e}} \sqrt{\sum_{V_e \in St(j)} \frac{1}{k_e}}} \tag{2}$$

Where St(i) is a set of node i and its neighbors, k_e is the degree of node e.

Because the Louvain algorithm has considered the weights of edges, we introduce weights into (2), improved similarity is defined as follows:

$$S(i, j) = \frac{\sum_{V_e \in St(i) \cap St(j)} \frac{1}{W_e}}{\sqrt{\sum_{V_e \in St(i)} \frac{1}{W_e}} \sqrt{\sum_{V_e \in St(j)} \frac{1}{W_e}}} \tag{3}$$

Where W_e is sum of the weights of the edges attached to node e.

2.2. Modularity Function Based on Node Similarity

Modularity is an evaluation criterion that measures the quality of community detection proposed by Newman, it is one of the most widely used methods at present. Modularity function is defined as follows in a weighted network:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \partial(c_i, c_j) \tag{4}$$

Where $A_{i,j}$ is the weight of the edge between node i and j, k_i is sum of the weights of the edges attached to node i, C_i is the community which i belongs to, m is all edges' weights, $\partial(c_i, c_j)$ is 1 if $c_i = c_j$ and 0 otherwise. The greater the value of Q is, the better the result of community detection is.

Reference [13] has put forward that we can measure if a community is reasonable from two angles: one is tightness between nodes within the community, the other is sparseness between nodes in different communities. Modularity belongs to the former. Formula (4) calculates modularity according to the weights of edges which can help judge whether two nodes belong to the same community. However, there still remains a lack of enough information to support the judgement. Node similarity, different from above, takes not only nodes themselves, but also their common neighbors into account. Therefore, it has enough information to judge if two nodes can be divided into the same community, thereby improving the accuracy of community detection.

We have analyzed the advantages of modularity based on node similarity, now we introduce similarity into modularity, it is defined as follows:

$$Q = \frac{1}{2TS} \sum_{i,j} \left[S_{i,j} - \frac{DS_i \cdot DS_j}{2TS} \right] \partial(C_i C_j) \tag{5}$$

Where TS is sum of similarity of all nodes, $S_{i,j}$ is the similarity between node i and j, DS_i is sum of similarity of i between its directly connected neighbors.

The gain of modularity ΔQ of moving i into a community C can be computed by:

$$\Delta Q = \left[\frac{\sum_{in} + s_{i,in}}{2TS} - \left(\frac{\sum_{tot} + s_i}{2TS} \right)^2 \right] - \left[\frac{\sum_{in}}{2TS} - \left(\frac{\sum_{tot}}{2TS} \right)^2 - \left(\frac{s_i}{2TS} \right)^2 \right] \tag{6}$$

Where $\sum in$ is sum of similarity of nodes that are directly connected to each other in C, $\sum tot$ is sum of similarity between nodes in C and their neighbors, s_i is sum of similarity between node i and its neighbors, $s_{i,in}$ is sum of similarity between i and its neighbors that are in C.

2.3. SHC Algorithm

SHC algorithm first introduces node similarity and makes an appropriate modification, then replaces weights in the modularity function with similarity. The whole algorithm is divided into three steps:

Step1 For an initial network of N nodes, compute the similarity between nodes that are

directly connected.

Step2 Treat each node as a separate community, then find all neighbors for each node, then move each node into the communities where its neighbors are in and compute the gain of modularity ΔQ, if ΔQ>0, then add the node into the community where ΔQ is biggest.

Step3 After step2, we get a preliminary partition of the network. Then we treat each community as a super node, and an edge represents the sum of similarity between two communities. We repeat these steps until there is no modularity gaining.

The pseudo code of the algorithm is as follows:

Input: network graph G(V，E) in adjacency matrix format.

Output: the number of communities and nodes in them.

1) For each {u,e}∈E
2) Compute node similarity S(i，j)
3) End for
4) Initialize each node in G as a community and compute modularity, denote it as Q_0
5) $Q_1=Q_0$
6) For i=1 to n
7) For j=1 to length(neighbour(i)) /*j is neighbor of i*/
8) Compute ΔQ of moving i to the community where j is in, taking the maximum value ΔQ_{max}
9) If $\Delta Q_{max}>0$
10) Move i to the community where j is in
11) End if
12) End for
13) End for
14) Compute modularity at the moment and store it into Q_0
15) If $Q_0>Q_1$，goto step5
16) Treat each community as a super node and repeat steps above
17) Return the result

3. Experimental Results

In order to evaluate the accuracy of the algorithm, we adopt the Normalised Mutual Information (NMI) which is an evaluation index of calculation accuracy in community detection algorithm proposed by the reference[13].NMI is a common evaluation index of calculation accuracy in community detection algorithm, many references[14][16][19] have adopted it. The value of NMI ranges from 0 to 1. The greater the value is, the more similar two communities are.

3.1. NCAA College-Football Network

The first application is on NCAA College-football network which is about American football games between Division IA colleges during regular season fall 2000.

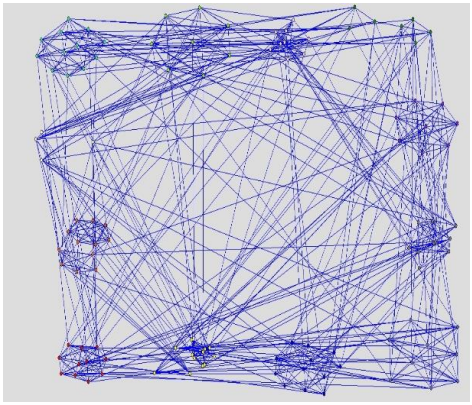


Figure 1. Original Community Structure

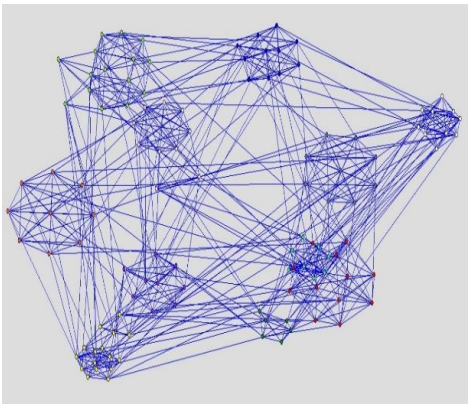


Figure 2. Detecting Result with SHC

As can be seen from Figure 2, SHC algorithm divides the football network into 11 communities, and the result is very close to the correct partition. For comparison, we also apply it with the Newman fast algorithm and the Louvain algorithm, the results are shown in Table 1:

Table 1. Detecting Results on NCAA College-Football

Algorithm	NMI	Best Q	Clusters
Newman fast algorithm	0.7678	0.4952	8
Louvain algorithm	0.8638	0.6046	10
SHC algorithm	0.8782	0.7681	11

Best Q in Table 1 represents the maximum of modularity gained by the algorithms, Clusters represents the number of communities detected corresponding to maximum value of modularity. As can be seen from Table 1, the modularity gained by Newman fast algorithm is the smallest, followed by Louvain algorithm and SHC algorithm. As we know, the value of modularity reflects the tightness between nodes within the communities. Therefore, we get more reasonable community structure due to modularity function based on similarity adopted by SHC algorithm. From the perspective of view of NMI, the value of NMI get by SCH algorithm is the greatest, while the greater NMI is, the closer it is to the standard results, so the SHC algorithm gets the best detecting result

on football network. On the other hand, when it comes to the number of communities detected, it is 11 by the SHC algorithm and 8 by the Newman fast algorithm and 10 by the Louvain algorithm. Obviously, 11 communities detected by the SHC algorithm is the closest to the original partition which is 12 communities. The result shows that the value of modularity and NMI are positively correlated meaning that the value of NMI is greater when modularity is greater. Therefore, we have proved SHC algorithm has the best detecting result compared to other two algorithms on the football network from two aspects of modularity and NMI.

3.2. LFR Network

Proposed by reference [13], LFR is a synthetic network usually used as benchmark network in community detection algorithm. In this paper, the program of LFR network provides 8 parameters to change community structure, they are as follows: the total number of nodes (N), the average node degree (k), the maximum degree (k_{max}), the node power-law distribution index (α), community scale power-law index (β), community structure definition (γ), community scale minimum (C_{min}), community scale maximum (C_{max}).

First, we generate a random network, observing the running effect of each algorithm. The parameters are shown in Table 2.

Table 2. The Parameters of LFR

N	k	k_{max}	α	β	γ	C_{min}	C_{max}
214	8	30	2	2	0.56	11	30

The results of experiment are shown in Table 3:

Table 3. The Results of Experiment on LFR

Algorithm	NMI	Best Q	Clusters
Newman fast algorithm	0.7104	0.2857	10
Louvain algorithm	0.8453	0.4709	10
SHC algorithm	0.9027	0.5620	12

As can be seen from the results of Table 3, the value of NMI and modularity in SHC algorithm are the greatest among the three algorithms, indicating the detecting result of the SHC algorithm is the best. In fact the generated network of Table 2 has 13 communities, so 12 communities detected by SHC algorithm is very close to the correct results.

To test affection of community structure definition on the accuracy of the SHC algorithm, we generate several networks that have different community structure definition through changing γ, the parameters are shown in Table 4.

Table 4. The Parameters of LFR

N	k	k_{max}	α	T	C_{min}	C_{max}
128	16	16	1	1	32	32

We compare the SHC algorithm with the Newman fast algorithm and Louvain algorithm. The results are shown in Figure 3.

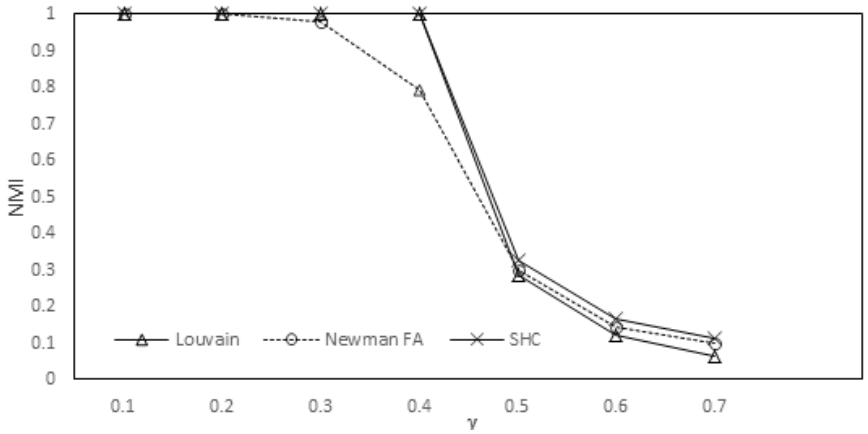


Figure 3. Variation Trend of NMI with Changing Γ

It can be seen that NMI value of three algorithms are 1 when $\gamma \leq 0.2$, meaning that the three algorithms can find correct community structure very well when community structure is obvious. When $0.2 < \gamma \leq 0.3$, NMI value of SHC algorithm and Louvain algorithm are still 1 while that of Newman fast algorithm is slightly less than 1. It shows that the detecting result of Newman fast algorithm is a little different from the standard result. When $0.3 < \gamma \leq 0.4$, NMI's value of the SHC algorithm and Louvain algorithm are 1 while that of Newman fast algorithm decreases obviously, indicating that there are some difference of detecting results between Newman fast algorithm and the correct result but the major is right. When $\gamma > 0.4$, NMI's value of the three algorithms decline sharply when γ increases, indicating that the detecting results become worse and worse when the community structure is more and more blurred, the SHC algorithm is higher than the other two algorithms among them. In a whole, the detecting accuracy of SHC algorithm is rather high when community structure is obvious, when community structure is not obvious, the detecting accuracy of SHC algorithm is higher than the Newman fast algorithm and Louvain algorithm.

4. Conclusion

This article analyzed a hierarchical clustering algorithm called Louvain algorithm. It continuously extracts community structure through optimizing modularity. Louvain has a very fast speed. However, because the modularity function of Louvain algorithm only considers link information between nodes when computing modularity, ignoring the surrounding neighbor nodes, resulting in decreased tightness between nodes in the same community. In order to solve this problem, we improved the modularity function with node similarity and proposed SHC algorithm. Then conducted experiments on the NCAA College-football network and LFR synthetic network, and compared it with the Newman fast algorithm and Louvain algorithm. The results revealed the NMI's value of SHC algorithm was the highest, proving that the accuracy of detecting results of SHC algorithm is much high.

Acknowledgement

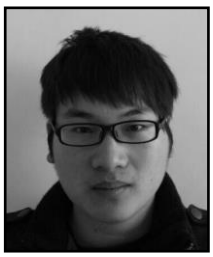
This work was supported by the National Natural Science Foundation of China (No.61402482), China Postdoctoral Science Foundation (No.2015T80555), Jiangsu Planned Projects for Postdoctoral Research Funds (No.1501012A), the Fundamental Research Funds for the Central Universities (No.2014QNB23).

Reference

[1] S. Fortunato and C. Castellano, arXiv:0712.2716, (**2007**).
[2] B. W. Kemighan and S. Lin, "An efficient heuristic procedure for portioning graphs", Bell System Technical Journal, vol. 49, no. 2, **(1970)**, pp. 291-307.
[3] Y. Liu, T. Yang, L. Fu and J. Liu, "Community Detection in Networks Based on Information Bottleneck Clustering", Journal of Computational Information Systems, vol. 11, no. 2, **(2015)**, pp. 693-700.
[4] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", PANS, vol. 99, no. 12, **(2002)**, pp. 7821-7826.
[5] M. E. J. Newman, "Fast algorithm for detecting community structure in networks", Phys. Rev. E, vol. 69, no. 6, **(2004)**, pp. 66-133.
[6] V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communitiesin large networks", Journal of Statistical Mechanics: Theory and Experiment, **(2008)**, pp. 10008.
[7] M. Girvan and M. E. J. Newman, "Finding and evaluating community structure in networks", phys. Rev. E, vol. 69, no. 2, **(2004)**, pp. 026113.
[8] S. Fortunato and M. B.´elemy, Proc. Natl. Acad. Sci. USA, vol. 104, no. 36, **(2007)**.
[9] S. Fortunato, "Community detection in graphs", Physics Reports, vol. 486, no. 3, **(2010)**, pp. 75-174.
[10] L. Min, Z, Liu, X. Tang and S. Liu, "A Local Community Detection Algorithm Based on the Weakening of Interference Nodes", Journal of Computational Information Systems, vol. 10, no. 19, **(2014)**, pp. 8295-8302.
[11] X. Xu, N. Yuruk and Z. Feng, "SCAN: a structural clustering algorithm for networks", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2007)**, pp. 824-833.
[12] X. Liu and D. Yi, "Complex Network Community Detection by Local Similarity", Acta Automatica Sinica, vol. 37, no. 12, **(2011)**, pp. 1520.
[13] J. Leskovec, K. J. Lang and M. Mahoney, "Empirical comparison of algorithms for network community detection", Proceedings of the 19th international conference on World wide web, vol. 6, **(2010)**, pp. 313-640.
[14] Y. R. Lin, Y. Chi and S. Zhu, "Facetnet: a framework for analyzing communities and their evolutions in dynamic networks", Proceedings of the 17th international conference on World Wide Web, **(2008)**, pp. 685-694.
[15] A. Lancichinetti, S. Fortunato and F. Radicchi, "Benchmark graphs for testing community detection algorithms", Physical Review E, vol. 78, no. 4, **(2008)**, pp. 46-110.
[16] K. Yu, S. Yu and V. Tresp, "Soft clustering on graphs", Advances in neural information processing systems, vol. 18, **(2006)**, pp. 1553.
[17] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis", Physical review E, vol. 80, no. 5, **(2009)**, pp. 056117.
[18] N. P. Nguyen, T. N. Dinh and Y. Xuan, "Adaptive algorithms for detecting community structure in dynamic social networks", INFOCOM, 2011 Proceedings IEEE, **(2011)**, pp. 2282-2290.
[19] J. Guo, A. Wei, J. Lv and Z. Liu, "An Improved Algorithm for Detecting Community Structure Based on Node Similarity", Journal of Computational Information Systems, vol. 10, no. 9, **(2014)**, p. 3805–3813.

Authors

Jingke Xi, He is an associate professor of College of Computer Science and Technology, China University of Mining and Technology. He received his Ph.D. degree at China University of Mining Technology in 2012; His research interests include community detection and data mining.



Wenwei Zhan, He is a postgraduate of College of Computer Science and Technology, China University of Mining and Technology. His research interests include data mining and social network analysis.

Zhixiao Wang, He was born in 1979 and received the Ph.D. degree in the Department of Computer Science and Engineering at Tongji University in 2011. He is an associate professor in the School of Computer Science and Technology, China University of Mining Technology. He has published more than 20 papers in international conferences and journals. His research interests include field theory application.

