Improving Louvain Algorithm for Community Detection

Bin Hu[1, a], Wenmin Li[1, b], Xuesong Huo[2], Ye Liang[3], Minghui Gao[3], Pei Pei[2]

[1]Beijing University of Posts and Telecommunications, Beijing, China;

[2]State Gird Jiangsu Electric Power Company, Jiangsu, China;

[3]Kedong electric power control system Co.Ltd, NARI Group Corepateion, Beijing, China.

[a]karas-hubin@foxmail.com, [b]liwenmin02@outlook.com

Abstract. Community is one of the important characteristics of reality network, which can effectively reflect the inner information of network and the relation among nodes. For the division of the community there already had many effective algorithms, the Louvain algorithm based modularity is a more popular community discovery algorithm because it can divide network into different hierarchical community structure quickly and efficiently. But, with the increasing size of network, the Louvain algorithm still has a serious problem that has relatively high time complexity in handling massive data. Faced with this situation, in this paper we ensure the merit of the Louvain algorithm and combine with the LPA algorithm which has advantage of effectiveness, proposing an improved algorithm integrating the Louvain algorithm with the LPA algorithm. Through later experiments, the improved algorithm can obviously decrease time complexity, reduce execution time, and ensure the result accuracy compared to original Louvain algorithm.

Keywords: Community detection; Louvain; LPA; Modularity; Time complexity.

1. Introduction

Community detection, it can be described as a behavior to explore cluster or society in the network. In a complex network consisting of points and edges, the connection between the two nodes within the same community is very tight, and the connection between the two communities is relatively sparse. For the feature, it is very significant for real network to discover community in all fields. In social network field, which can be divided into various social communities based on different interests and backgrounds for accurate recommendation. In biology field, the propagation network consists of disease and virus, which can be researched and analyzed to find the key community or key node where some people be infected easily, to predict the propagation path so that disease can be promptly controlled. In the field of web security, we can execute anomaly detection for network host divided into different communities, then take control of the communities which containing of anomaly flows in time. In summary, community detection plays an important role in studying properties of complex networks.

In recent years, the field of community detection has attracted more and more attention, its theoretical value and practical value are acknowledged by researchers, and about it more algorithms have been proposed. From algorithm ideas, community detection algorithms have two main types, 1. Clustering algorithm, which adds the edges of network graph. 2. Classification algorithm, which removes the edges of network graph. From algorithm result, the algorithms of community detection can be divided into non-overlapping community and overlapping community based on whether a node belong to multiple communities. From algorithm method, Raghavan [1] proposed LPA (Label Propagation Algorithm) algorithm based on the idea that edges of network graph sometimes represents the dissemination of information [2]. Proposed GN algorithm to split community. Modularity for quality evaluation was proposed by Newman and Girvan [3] in 2004. Shiokawa [4] used Spectral clustering algorithm based on modularity to divide community. Peng [5] thought that the partly important nodes of graph can decide the overall framework of community, therefore adopted K-core algorithm and so on. All above algorithms have advantages and shortages, for example, GN algorithm is too complex to use in large scale graph. LPA algorithm has linear time complexity but lacks of accuracy.

With the establishment of a variety of big data computing platforms, machine bottleneck had already no longer exist, and algorithm complexity becomes the key factor that restrict the performance of community detection in real networks. So, this paper focuses on how to decrease time complexity and reduce execution time. Taking into account the time, network size and other factors, the Louvain algorithm proposed by Blondel el al [6] stands out from other algorithms because it has the advantages of performance. It is an important goal of some researchers to improve the Louvain algorithm. Such as [7] improved the accuracy by adding the 'uncoarsening' phase, [8] implemented parallel algorithm to improve the efficiency. This paper proposes a hybrid community detection algorithm which is combined with the Louvain algorithm and the LPA algorithm to improve the efficiency of algorithm. From the later experiment results, the hybrid algorithm can effectively improve the performance of the original algorithm.

2. Related work

In this part, we mainly introduce the existing work proposed by the previous researchers about our algorithm. It introduces the concept of modularity, describes the Louvain algorithm's execution process and main idea, and meanwhile describes the characteristics of the LPA algorithm.

2.1 Modularity

Modularity is proposed by Newman [4], which is a commonly used method to measure the stability of the network community. It can be defined that networks of high modularity have dense connections among the nodes within same community and sparse connections among the nodes in different communities. So, the higher value the modularity has, the more stable the network. The formula is as follow:

$$\Theta = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \qquad (1)$$

Where A_{ij} represents the weight of the edge between i and j. m=$\frac{1}{2}\sum_{i,j} A_{ij}$ is the sum of the weights of the edges of the graph. c_i is the community which vertex i is assigned to and k_i is the sum of the weights of the edges attached to vertex i. If c_i=c_j, the δ-function (c_i,c_j) is equal to 1, otherwise the function value is 0. To Simplify formula (1), we can get a formula (2) as follow:

$$\Theta = \sum_c \frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 \qquad (2)$$

Where $\sum$ in represents the sum of inner edges of a community, $\sum$ tot is the sum of all edges of a community.

2.2 Louvain Algorithm

Louvain algorithm was proposed by Vincent D.Blondel [6] in 2008, which is a heuristic algorithm based modularity optimization. Louvain algorithm is an effective method to divide community and it is an iterative algorithm through repeatedly dividing communities to get the maximum modularity of entire network. The main steps of Louvain algorithm are as follows:

1) First, we assign a different community to each node of the network, so in initial partition the number of the communities is as many as the nodes'.
2) For each node we consider its neighbors and we evaluate the value of modularity after removing one node from its community and placing it in one of its neighbor's community. The modularity can be calculated by formula If ΔQ is positive, the node stays in its original community, otherwise the node is placed in the updated community.

$$\Delta Q = [\frac{\sum_{in} + k_{i,in}}{2m} - (\frac{\sum_{tot} + k_i}{2m})^2] - [\frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 - (\frac{k_i}{2m})^2] \qquad (3)$$

3) Repeat 2) until communities of all nodes no longer change.
4) Construct a new Graph and each node represents a community partitioned by 3). Executing 2) and 3) continuously until gaining the maximum modularity value.

2.3 LPA Algorithm

Label Propagation Algorithm is a partial community partition algorithm, which based on label propagation. The algorithm procedures are listed as follow:

1) First: We assign a different community to each node of the network like the first step of the Louvain algorithm.
2) Second: For each node, we changed its community label according to the maximum number of its neighbor labels which belong to same community.
3) Third: If every node has a community label that the maximum number of its neighbors have, and then stop the algorithm, else continue second step.

The biggest advantage of LPA algorithm is that it has a relatively simple process compared with modularity optimization and runs very fast but effect is unstable.

3. Algorithm Optimization

From the above description, the Louvain algorithm needs constant iteration to find the best community for each node in step 2, and the process spends a lot of time in large networks. Meanwhile LPA algorithm has liner time complexity and fast execution process, but the result quality after divided community is not good. So, the improved algorithm proposed by this paper can be summarized that the network can be divided into two sub-graphs according to the key nodes which have more edges and edge nodes which have fewer edges. The key nodes use Louvain algorithm to divide network community for better quality, and the edge nodes use LPA algorithm for less execution time. So, the improved algorithm can decrease time complexity for executing fast, while ensures the detection quality. It is different from the previous pruning algorithm which only apply once in the algorithm begin, the LPA algorithm is applied to every iterative procedure of the Louvain algorithm. We called it as Louvain-LPA algorithm and it has five stages as follows:

1) Label the nodes that the number of edge is less than K in the graph G, and compute the relate information for each node and its neighbors.
2) Execute 2pass and 3pass of the Louvain algorithm on the unlabeled nodes.
3) Execute LPA algorithm on the labeled nodes after the procedure of 2).
4) Execute 4pass of the Louvain algorithm.
5) Execute iteratively the procedure of 1), 2), 3) and 4) until satisfied the condition that gets the maximum modularity.

This improvement mainly reduces the nodes handled by the Louvain algorithm which consumes much time, then uses the LPA algorithm which consumes less time to compute remain nodes for nodes' community. This method applies to every time for the iterative procedure of the Louvain algorithm. It can effectively decrease time complexity of algorithm, and accelerate the process of community detection.

Through the above parts, it is obvious problem that how to determine the value of K, and the different K value can make a difference on execution time and effect of the algorithm. The principle about K value should be smaller influence accuracy, meanwhile promote efficiency greatly, so the K value can be dynamic value according to the actual network node distribution. From the graphics, we can know that graph has an important property, which is degree. Degree presents the sum of income edges and outer edges of a node, and the average degree is as D=e/v, where e is the total number of edges of the graph, and v is the total number of vertexes of the graph. In this paper, we define the key value as K=D.

4. Experiment

In this part, in order to verify the effect of the improved algorithm, we decided to do experiments respectively from the divided community quality, execution time of algorithm, the influence of K value. These experiments mainly show the improved Louvain algorithm compared to original Louvain on efficiency improvement, meanwhile verify the reliability of the algorithm. The experimental details as follows.

Table 1 Real network datasets

Dataset	type			
	Football	***Collaboration***	***Friendship***	***Amazon***
Vertex	114	9875	58228	334863
Edge	613	25973	214078	925872
Average Degree	5.3(5)	2.6(3)	3.6(4)	2.7(3)

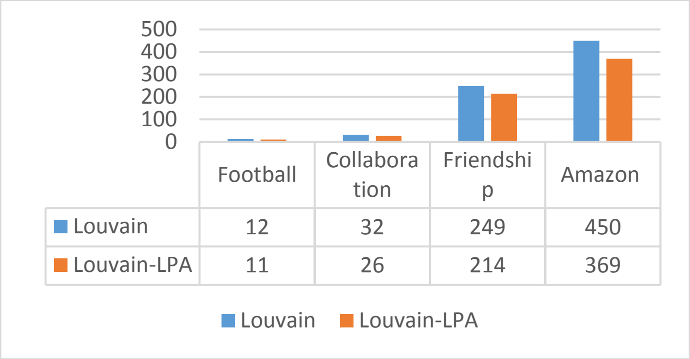
Our experiments use the VMware of Xen HVM domU which has 4GB of RAM and Intel(R) Xeon(R) CPU E5-2609 of 2.40GHz. The experiment datasets come from the Internet where many real network datasets is available. For the experimental accuracy in different network structure, we choose four kinds of networks shown in Table 1 to evaluate our improved algorithms. The datasets is as follows: Football, a network of American football games between Division IA colleges during regular season fall 2000 and the average degree is 5.3. Friendship, a location based friendship network consisting of 58,228 nodes and 214,078 edges, which the average degree is 3.6. Collaboration, a collaboration network of 9,875 authors working on the theory of high energy physics and the average degree is 2.6. Amazon, a network consists of 334,863 edges, 925,872 nodes and the average degree is 2.7, which each node represents a product on the Amazon website

4.1 Time Consumption

In order to compare the execution time difference between the improved Louvain algorithm and the original Louvain algorithm, we respectively run the Louvain algorithm and the Louvain-LPA algorithm in four kinds of datasets of real network, and the experiment result is shown in Fig.1. The chart displays the execution time of two algorithms in different datasets, in where the vertical axis is the execution time, and the horizontal axis is the dataset kind. From this Fig.1, we can see that the Louvain algorithm respectively spent 12seconds, 32seconds, 249seconds, 450seconds on four datasets, while the Louvain-LPA algorithm spent 11seconds, 26seconds, 214seconds, and 369seconds. It is obvious that the more nodes and edges, the more much execution time of the Louvain algorithm and the Louvain-LPA algorithm. We can observe that Louvain-LPA algorithm in each dataset relative to the Louvain algorithm has better performance through less execution time. We can also find that with the increase of nodes and the decrease of the average degree of network, the Louvain-LPA algorithm performance is more outstanding, and its execution time compared to the Louvain algorithm respectively decreases by 9.09, 23.07%, 16.35%, and 21.95%. This experiment verifies that compared to the original Louvain algorithm, the Louvain-LPA algorithm has a lower time consumption in the same dataset.

4.2 Modularity similarity

The Louvain algorithm is a heuristic algorithm based on modularity, so modularity is the important standard to measure quality of the Louvain algorithm. In this experiment, we mainly evaluate the Louvain-LPA algorithm relative to the Louvain algorithm in the aspect of modularity influence. After running respectively two algorithms in four kinds of complex network dataset consisting of Football, Collaboration, Friendship, Amazon, and computing the modularity of community detection. The result is such as Fig.2. From the Fig.2, the modularity of the Louvain algorithm respectively are 0.50, 0.485, 0.38, 0.65, and the modularity of the Louvain-LPA algorithm are 0.48, 0.487, 0.43, and 0.68. The similarities respectively are 96%, 99%, 88%, and 95%. we can know the result that the Louvain-LPA algorithm and the Louvain algorithm there exists a positive correlation, and the modularity of two algorithm results has a high similarity, so it can prove that the Louvain-LPA algorithm only has a slight influence about the accuracy. We can also find that the Louvain-LPA has a better performance in network which consists of more nodes and more edges, such as Amazon dataset.



	Football	Collaboration	Friendship	Amazon
■ Louvain	12	32	249	450
■ Louvain-LPA	11	26	214	369

Fig.1. Algorithm running time

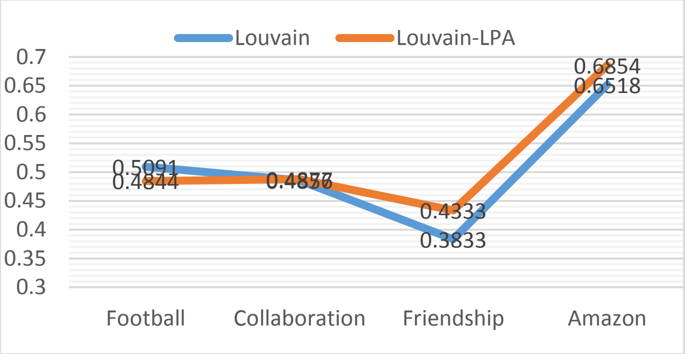


Fig.2. Modularity of contrast

4.3 K-Value influence

In the improved algorithm of Louvain, the selection of K value has a significant impact to both efficiency and accuracy of the algorithm. K value determines the proportion of nodes handled respectively by the Louvain algorithm and the LPA algorithm. When k is bigger, the Louvain algorithm handles fewer nodes, and the LPA algorithm handles more nodes. In the previous experiments, the K value is gained as the average degree. So, we have adopted this experiment to verify the effect of K value. In the experiment, Amazon dataset which have most nodes was chosen as experimental dataset, then computing the execution times and modularity of the Louvain-LPA algorithm when K value respectively is 1, 3, 5, 7, 9, 12, and 15. The experimental result is as Fig.3, where it can be seen that maximum modularity is 0.68 when K is between 3 and 5, and the maximum execution time is 495seconds when K value equals 1. With the increment of K value, the execution time is obvious decline, the modularity is also from stable to rapid decline, and there was a positive correlation between them. From the III chapter, we know the average degree of Amazon is 2.7(3) and it is obvious that when K value equal between 5 and 7, the algorithm effect is best because the modularity/time ratio is highest. So we can make a conclusion that the Louvain algorithm determines the accuracy of the algorithm, and the LPA algorithm determines the efficiency of the algorithm. It is important that a suitable K value will determine the quality of improved algorithm.

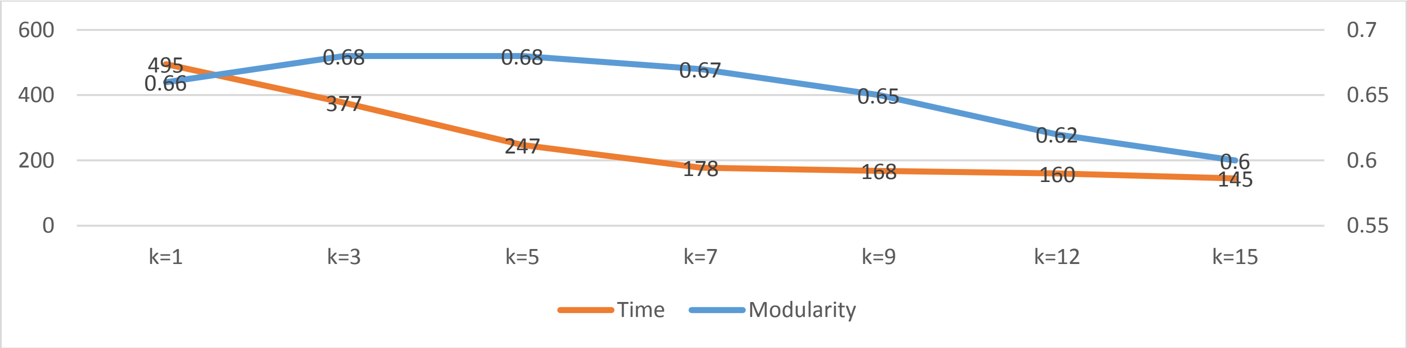


Fig.3. the influence of K value

5. Conclusion

Community detection has a great significance in the application of complex network, and the Louvain algorithm which is based on hierarchy due to its efficiency and good accuracy has been widely applied. This paper first points out the shortage of the original Louvain algorithm, then proposes the improved method in the aspect of time complexity. Through combining the LPA algorithm and the Louvain algorithm, we have introduced a class of hybrid algorithm called Louvain-LPA algorithm to identify communities in network. The Louvain-LPA algorithm reduces the time consumption in the iterative phase. The following experiments prove that this method has achieved good effect about algorithm execution time in the real network datasets, so this improvement of the Louvain algorithm is effective.

Acknowledgments

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

References

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

[2] 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.

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

[4] Shiokawa H, Fujiwara Y, Onizuka M. Fast Algorithm for Modularity-Based Graph Clustering [C]//AAAI. 2013.

[5] Peng C, Kolda T G, Pinar A. Accelerating community detection by using k-core subgraphs [J]. arXiv preprint arXiv:1403.2226, 2014.

[6] Blondel V D, Guillaume J L, Lambiotte R, et al. fast unfolding of communities in large networks [J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

[7] Gach O, Hao J K. Improving the Louvain Algorithm for Community Detection with Modularity Maximization [C]//Artificial Evolution. Springer International Publishing, 2013: 145-156.

[8] Que X, Checconi F, Petrini F, et al. Scalable Community Detection with the Louvain Algorithm [C]//Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International. IEEE, 2015: 28-37.

[9] Li W, Jiang W, Hua-Wei S, et al. An improvement of the fast uncovering community algorithm [J]. Chinese Physics B, 2013, 22(10): 108903.

[10] Shiokawa H, Fujiwara Y, Onizuka M. Fast Algorithm for Modularity-Based Graph Clustering [C]//AAAI. 2013.