http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (3): 384-387, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.384.387

Community Detection via Improved Genetic Algorithm in Complex Network

Shangguang Wang, Hua Zou, Qibo Sun, Xilu Zhu and Fangchun Yang
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications,
Beijing 100876, China

Abstract: In this paper, we propose an improved genetic algorithm for community detection in complex networks. The basic idea of the improved genetic algorithm is to modify crossover operators which are more suitable for community detection and then design a heuristic mutation operator based on local modularity to avoid the blindness of random flip. The experimental results show that the improved genetic algorithm can reduce the selection pressure.

Key words: Community detection, complex network, genetic algorithm

INTRODUCTION

There have been numerous and various complex networks with the development of science, technology and human society. Recently, the other characteristic of the complex networks is the community structure, causes lots of focus. In general, a community can be considered that the nodes in the same community have an intensive connection with each other, but the nodes from different community have a sparse connection. Hence, the father the distance of two nodes, the greater the probability they belong to different subcommunity is. The community mining has been successfully applied into the area including the identification of terrorist organizations, web community detection, web search engine and so on.

Due to the vast application prospects and theory of value, lots of job has been done and numerous community mining algorithms have been proposed. These algorithms can be classified into two types: one is based on heuristic algorithms; the other is based on the optimization algorithms. The former algorithms include the famous Girvan-Newman algorithm (Girvan and Newman, 2002), detecting network communities by propagating labels (Raghavan *et al*., 2007), the propinquity dynamics algorithm (Zhang *et al*., 2009). However, the latter convert the community detection algorithm to the optimization problems by using an objective function to search the communities in the complex network. However, the optimization problem is an NP hard puzzle, it's practice infeasible to carry out an exhaustive search of all possible division. Therefore, the genetic algorithm which is one of the combination optimization algorithms has been introduced.

Genetic algorithm is the effective method to solve the combination optimization problem, which guarantees the quality of the solution with low complexity time. Liu *et al*. (2007) proposed a community detection algorithm based on genetic algorithm where the crossover operator isn't adopted. All these algorithms mentioned above use the simple genetic algorithm might suffer the same problems that the searching ability is not so good and it is easy to trap in the local-optimization. Additionally, there exist some problems in their utilization. In some cases this makes it impossible to run an application efficiently using a single machine. Hence, the parallel implementation of genetic algorithm can provide gains in terms of performance and can search in parallel different subspace of the search space, which makes it less likely to by low-quality subspace.

In order to overcome these problems, in this study, we propose an improved genetic algorithm for community detection in complex networks. In the improved genetic algorithm, we first improved its crossover operators to make it more suitable for community detection and then to avoid the blindness of random flip, we design a heuristic mutation operator based on local modularity. The experimental results show that the improved genetic algorithm can effectively avoid the selection pressure.

COMMUNITY DETECTION

Genetic algorithm is a computation mode which simulates the process of natural evaluation. It has been widely used in combinatorial optimization, system control, machine learning and data mining fields. Thus, it also is

Corresponding Author: Shangguang Wang, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

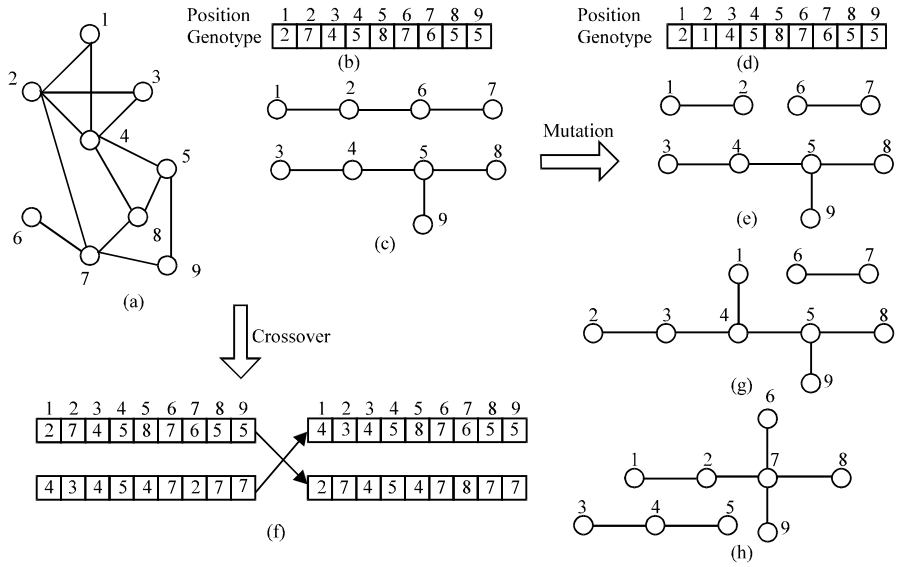

Fig. 1: Genetic operator

suitable for solving the community detection. The evolutionary operator for detecting community can be depicted as shown in Fig. 1.

Figure 1a shows that a network modeled as a graph. Figure 1b shows that the locus-based representation of a genotype. Figure 1c shows that the graph-based structure of the genotype. Figure 1d shows that the graph-based structure of the genotype after mutation, Fig. 1f showing the crossovers, Fig. 1e shows the mutations and Fig. 1g and h show that the graph-based structure of the genotype after crossover.

In the genetic algorithm, each chromosome uses the locus-based adjacency as the encoding scheme. Therefore, a node j assigned to the i-th gene is interpreted as a link between the node i and node j. The genetic operator adopts the traditional method, such as uniform crossover and a bit-flip mutation. While this process must ensure the ith gene value is a neighbor of ith gene. After one times evolutionary iteration, a chromosome can divide the graph into several communities which can be denoted by the maximum connecting graph in the chromosome. The partition example is shown at Fig. 1c. The partition result can be assessed by using the modularity Q proposed by Newman and Girvan (2004) as follow:

$$Q(C) = \sum_{c \in C} \left[\frac{|E(c)|}{m} - \left(\frac{\sum_{v \in c} \deg(v)^2}{2m} \right) \right] \quad (1)$$

where, C denotes the community result E (c) denotes the internal edge set in a community (v) denotes the degree of node v, m denotes all edge in a community. To obtain the maximum Q value, the value of the first part in equation should be as large as possible and the second part of equation should be as small as possible.

The traditional genetic algorithm has some shortcomings applied into the community detection, such as premature, low convergence speed. To solve these problems, a improved genetic algorithm is proposed. The basic idea is to modify crossover operator and propose a new mutation operator.

Modified crossover operator: In genetic algorithms, crossover is essential operator which enhances the convergence rate. It combines two or more parents to reproduce new children, then, one of these children may hopefully collect all good features that exist in his parents. However, the traditional crossover operator is not suitable for the community detection since the building blocks which might not be sequence. Thus, the crossover operator might destroy the building blocks pile up.

According to the result of community detection, the building blocks hiding in the community might not be a continuous sequence. Apparently, The traditional crossover contribution may not so good as the method which pick out a section of largest connected subgraph to crossover.

The definition of the building blocks can be considered as a short, low order, highly fit schemata. To satisfy the condition, we set s constant to limit the length of gene sequence to swap. Therefore, the parents involved in crossover operator mainly swap a fixed length

of a connected subgraph in a chromosome. The time complexity of the crossover operator for a population is O (n).

Although, the crossover contributes to the convergence, it often lead to premature convergence without mutation operator. To satisfy the effectively searching community in the complex network, we propose a heuristic mutation operator to avoid genetic drift.

A new mutation operator: The crossover operator ensures that excellent scheme will be inherited by the descendants. However, the local convergence must rely on the mutation. Although, the mutation adopts the strategy which randomly selects one gene value to flip. the blind mutation is inevitable to lead the genetic drift and lower the convergence speed. To speed up the local convergence rate, we propose a heuristic mutation operator in terms of the local modularity definition.

Feng *et al*. (2006) proposed the measure called modularity M for local community evaluation. It directly compare the ratio of internal and external edges:

$$M = \frac{edges_{internal}}{edges_{external}} \quad (2)$$

where, $edges_{internal}$ means two endnodes are both in the same community and the $edges_{external}$ represents two of them belonging to the different community. Therefore, by integrating with the local modularity, the mutation operator might avoid the blind search and the mutated gene value is chosen from its neighbors which has maximum edges in the local community. The time-complex cost mainly concentrate in the procedure of selecting the maximum ediges.

Therefore, by integrating with the local modularity, the mutation operator might avoid the blind search and the mutated gene value is chosen from its neighbors which has maximum edges in the local community.

EXPERIMENTS

Here, we show the results obtained using the improved genetic algorithm to detect communities on undirected networks. The cases study used in our experiments for community detection are political blogosphere network. The cases study used in our experiments are political blogosphere network based on incoming and outgoing links and posts around the time of the 2004 presidential election with 19025 links where the topology of the Western States Power Grid of the United States with 4941 nodes and 6594 links and the Zachary krate club with 34 nodes and 78 edges.

Table 1: The speedup experiment

Type	R=1	R=5	R=10
karate	21342	22938	23840
polblog	70001	59992	57321
power	247304	100024	79259

Table 1 illustrates the running time in different reducer number. The speedup for karate club network is not so good as the standard genetic algorithm and the time consumption is growing bigger with the reducer number increasing. This is also why the speedup for political blog network obtains is not apparent. From the power data, the time consumption reduces to 1/3 of the original when the reduce number is less than 10.

CONCLUSIONS

In this paper, we propose an improved genetic algorithm for community detection. We first modified its crossover operators to make it more suitable for community detection and then we design a heuristic mutation operator based on local modularity to avoid the blindness of random flip. Simulate results confirm that the improved genetic algorithm can effectively perform community detection and avoid the selection pressure to destroy the diversity.

ACKNOWLEDGMENT

Thanks for 973(2009CB320406); 863(2008AA01A317) and "HeGaoJi" (2009ZX01039-001-002-01).

REFERENCES

Feng, L., J.Z. Wang and E. Promislow, 2006. Exploring local community structures in large networks. Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, December 8-22, Hong Kong, pp: 233-239.

Girvan, M. and M.E.J. Newman, 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA., 99: 7821-7826.

Liu, X., D. Li, S. Wang and Z. Tao, 2007. Effective algorithm for detecting community structure in complex networks based on GA and clustering. Proceedings of the 7th International Conference on Computational Science, Part II, May 27-30, 2007, Beijing, China, pp: 657-664.

Newman, M.E.J. and M. Girvan, 2004. Finding and evaluating community structure in networks. Phys. Rev. E-Stat., 69: 026113-026128.

Raghavan, U.N., R. Albert and S. Kumara, 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E, 76: 36-106.

Zhang, Y., J. Wang, Y. Wang and L. Zhou, 2009. Parallel community detection on large networks with propinquity dynamics. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 1, 2009, Paris, France, pp: 997-1006.