Community Detection in Social Graph Using Nature-Inspired Based Artificial Bee Colony Algorithm with Crossover and Mutation

Thet Thet Aung, Thi Thi Soe Nyunt, Pyae Pyae Win Cho
University of Computer Studies, Yangon
Myanmar
e-mail: thetthetaung@ucsy.edu.mm, thithi@ucsy.edu.mm, pyaepyaewincho@ucsy.edu.mm

***Abstract*—Many types of social network are modelled as graphs. Community detection has been an important research area in social graph analysis. Community detection can be viewed as an optimization problem. Nowadays, researchers use nature-inspired algorithms to solve optimization problem. Their goal is to find the optimal solution for a given problem. In this paper, nature-inspired based artificial bee colony algorithm with crossover and mutation is used to detect community in social graphs. GraphX is built as a library on the top of Spark by encoding graph as a collection of vertices and edges. Comparative studies describe that the proposed algorithm and other nature-inspired algorithms can effectively detect the community structure on real world social graphs as other traditional community detection algorithms.**

Keywords-artificial bee colony; community detection; graph; GraphX; spark; crossover; mutation and modularity

I. INTRODUCTION

Real social network can be represented as graph. Community detection has become one of the important research areas in social graph analysis. Each graph contains communities in which vertices in a community have dense connections and sparse connections to the other communities. The examples of communities in the graph are a group of consumers in similar interest, group of researcher interested in same topic and so on.

Many traditional community detection algorithms are implemented in igraph package. Some community detection (CD) algorithms based on hierarchical algorithm. Hierarchical CD algorithm divides or merges communities by considering nodes distance or similarity between them. Some algorithms are optimization based algorithm that divide the network based on a given criterion. In optimization approaches, an objective function is defined to measure the quality of result communities. CD algorithm efficiency depends on optimization algorithms, objective function and search strategy of algorithm.

Nature-Inspired algorithms are widely used in optimization problem. They have demonstrated their capacity to produce the acceptable solution in many real life complex problems as community detection. Generally, there is a set of candidate solution, which are numerically encodes and an objective function to be optimized on the solution space. They have improved the quality of communities found by heuristic search community detection algorithm. Genetic algorithm, evolutionary algorithm, particle swarm optimization algorithm and ant colony algorithm are mostly used algorithms to solve optimization problem.

Artificial bee colony algorithm is nature-inspired based meta-heuristic algorithm. It simulates the foraging behavior of honey bee. Many applications, modification and comparative studies on numerical optimization problem of artificial bee colony algorithm (ABC) get high performance and accurate results [1]. So, ABC is selected due to high performance of ABC on the numerical optimization problem. In this paper, it is used to solve community detection problem. Crossover and mutation are used within ABC algorithm to search new food source. Experiment results on real social graphs reflect that the proposed algorithm can detect communities effectively.

The rest of this paper is organized as follows. Section II presents the related works concern with community detection in social graph and III describes about basic theory. Section IV presents the proposed algorithm. Section V shows the experimental results of different CD algorithm on four social graphs and VI serves as conclusion and future work.

II. RELATED WORKS

Researchers proposed many community detection algorithms in many research fields over the year. All are efficient and effective in their way. Igraph community detection functions can be used to detect communities for undirected graph. For these algorithms typically the membership for the highest modularity value is returned. These algorithms can be tested on RStudio [2].

A parallel particle swarm optimization PSO based on Apache Spark is used for community detection in social network by Shanfeng Wang et al. Modularity density is used as the objective function. GraphX is also used to optimize modularity density parallel [3]. Evolutionary algorithm for community detection is used by Gema et al. using graph based information. Binary encoding was used where binary 1 used to denote the node belong to the community [4]. Ramadan babers et al. used Ant Lion Optimization Algorithm for community detection. They chose community fitness and modularity as quality functions [5].

Youcef Belkhiri et al. proposed bee swarm optimization for community detection in complex network. In this proposed algorithm, modularity is taken as objective function, k is used to create bees in a search area and also use taboo list to avoid cycles during the research process [6]. Ronghua Shang et al. improved genetic algorithm to detect community in complex network. Modularity and improved

genetic algorithm (MIGA) is proposed to solve community detection problem and used prior information of network [7].

In this paper, ABC algorithm with crossover and mutation is used to detect community in social graph.

III. BACKGROUND THEORY

Background knowledge about artificial bee colony algorithm and crossover and mutation are discussed in this section.

A. ABC Algorithm

Artificial bee colony algorithm is one of nature-inspired based algorithms. ABC algorithm contains food sources. Each food source has d dimensional vectors that are updated by the artificial bee [8]. The goal of bees is to find the better food sources which have optimal nectar amount.

Employed bees, onlooker bees and scout bees are used for search strategy. Employed bees generate new food sources v_i using neighborhood food source in the solutions. After that, they share their food sources information to onlooker bees. Onlooker bees choose the food sources depending on its nectar amount. If the nectar of a food source increases, the probability value of food source will be preferred by onlooker bees increases too. The new food source's nectar amount is more than the previous food source; bee updates its food source and forget the old one. If the food source is abandoned, the scout bee searches new random food source.

Three bees' processes are repeated until a termination criterion is satisfied. In ABC, exploitation is made by employed and onlooker bees and scout control exploration. ABC attempts to balance between exploitation (local search) and exploration (global search). The general structure of the ABC algorithm is given as follow.

Initialize population of solutions
Generate fitness of all solutions
Keep best solution in the population
REPEAT
    Employed Bees: Evaluate new solutions
    Onlooker Bees: Evaluate new solutions based on food source's probability value
    Scout Bees: if abandon solution exists, generate new random solution
    Memorize the best solution
UNTIL (Maximum Cycle Number)

B. Crossover and Mutation

Crossover and mutation is the most significant phase in genetic algorithm (GA). In GA, chromosome represents a candidate solution. Each chromosome can be divided into genes. Each genes represents a particular element of chromosome. For solution representation, different type of encoding schema is used such as binary encoding, integer-encoding, tree encoding and etc. When selecting parents' chromosomes for reproducing according to their fitness value, crossover and mutation operator take place. Crossover consists of exchanging genes between two chromosomes (parents) and forming new chromosomes (offsprings). Crossover occurs with some probability (crossover rate). If there is no need to take crossover, offspring is simply a copy of its parents. Mutation is performed by choosing a gene at randomly chosen locus and replacing that gene with another one. Mutation operator maintains diversification in the population ensuring that no single gene position keeps fixed value during the algorithm run.

IV. PROPOSED ALGORITHM FOR COMMUNITY DETECTION

In this section, artificial bee colony algorithm with crossover and mutation is proposed to solve community detection problem. In ABC the number of food sources represents the population of solutions. The good food source indicates a better solution to the given optimization problem. This algorithm is implemented on Spark framework and GraphX.

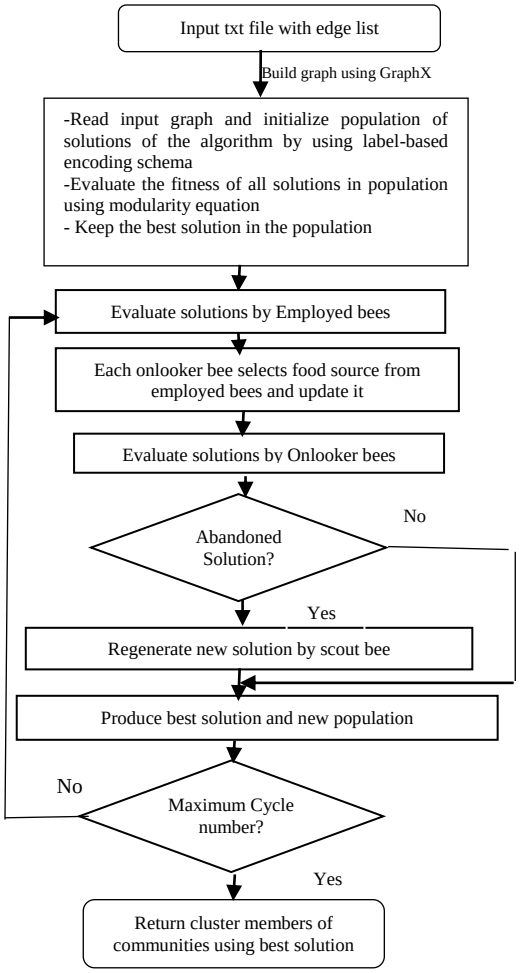

Figure 1. Flow diagram of the proposed algorithm.

GraphX represents graphs internally as a pair of vertex and edge collections. It built on Spark RDD abstraction. In the rest of this section, create graph using GraphX, the solution representation and initialization, fitness function and search strategy of proposed algorithm are also described.

A. *Creating Graph Using GraphX*

The input text file contains edge lists of social network. GraphX on Spark framework is used to create graph. The input txt file of sample network in figure 2 is 1 2,1 3, 2 3,3 4,4 5,4 6,4 7,5 6,7 8. Graph<VD, ED>, VertexRDD <Long> and EdgeRDD <Long> in graphX are used to create graph from txt file. Graph contains the set of vertices and edges in the given network. Sample graph gets 8 vertices and 9 edges. Then create internal graph **Map<Long,Set<Long>> internalGraph = new HashMap<Long,Set<Long>>()**. That internal graph contains vertices and its neighbor vertices list. The output of internal graph is

Node id : 4 neighbour ids 3, 5, 6, 7,
Node id : 6 neighbour ids 4, 5,
Node id : 8 neighbour ids 7,
Node id : 2 neighbour ids 1, 3,
Node id : 1 neighbour ids 2, 3,
Node id : 3 neighbour ids 1, 2, 4,
Node id : 7 neighbour ids 4, 8,
Node id : 5 neighbour ids 4, 6.

First line means that node 4 has links to node 3,5,6 and 7. It is used in many parts of the proposed algorithm. In initializing population of solutions and finding fitness function, it can be used.

B. *Representation And Initialization*

The input of community detection problem is a graph structure and the output is community member of each node in the graph. In graph G(V,E), V is the number of v vertices and E is set of e edges. Each food sources in the population has in the form $X_i=(x_{i1},x_{i2},x_{i3}...,x_{iv})$ where v is the number of vertices in the graph. For solution initialization, label-based representation is used. The value of each vertex index indicates to which the community label the vertex belongs. There are p number of population of solutions. After discussing about the representation, this paper presents about the initialize population.

Label-based encoding schema can automatically define the numbers of community in network. To get maximum numbers of community, start from first index vertex x_{i1}. $x_{i1}$and their neighbors are assigned same community label and then the number of community will increase. Then go to next index x_{i2}. If the second index has own community label, skip it. Second index vertex and their neighbors will be assigned same community label if the vertices don't have label value. This assignment will make until the last vertex. Finally, one food source and the maximum number of community is gotten. Based on the food source, initialize the population of solutions. For each food source, a vertex t is randomly chosen and assigned its community label to all of its neighbors. The random probability is 0.3 for each food source.

C. *Fitness Function*

Fitness function is main factor of swarm intelligence algorithm. This paper used modularity Q as a fitness function of the proposed algorithm. It does not need to know ground truth communities of networks. Modularity evaluate the quality of a partition of nodes in a graph as good communities. Modularity is the difference between the number of links in a group and the expected number of links in the same group of a comparable random graph.

$$Q = \frac{1}{2m}\sum_{ij}(a_{ij} - \frac{k_i\, k_j}{2m})\delta(c_i\,, c_j\,) \quad (1)$$

Where, k_i is the degree of vertex i and k_j is the degree of vertex j, c_i is the community label of node i, c_j is the community label of node j, a_{ij} =1 if node i and node j have link that value can be gotten from the internal graph and δ function is 1 if i and j have same community label, otherwise it equals to 0, m is the total number of edges in the graph [9].

D. *Search Strategy*

Employed bees search new food sources using the neighborhood food sources in the population. The new food source has more nectar amount than old food source. Crossover and mutation are used when evolving new food source in ABC algorithm. One-way crossover is used in crossover operation. Select a crossover point randomly. Following example shows crossover operation on a food source. For each food source, choose a neighbor food source in the population. The nectar amount of neighbor food source is higher than other food source in population.

Example: for figure 2 sample graph, randomly choose crossover node is 3 in original food source. Select other nodes with same community label with crossover node depend on original food source. In example, node 3's community id is 2. Then, choose other nodes that has same community id as node 3. Swap genes values in original food source with same genes indexes values in neighbor food source.

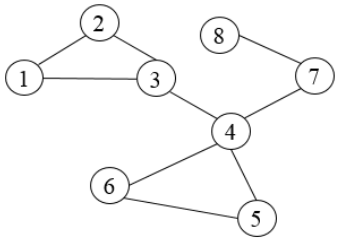

Figure 2. Sample graph.

Original Food source

1	2	3	4	5	6	7	8
1	1	2	1	2	1	3	3

Neighbor Food Source

1	2	3	4	5	6	7	8
1	1	1	2	2	2	3	3

Child food source 1

1	2	3	4	5	6	7	8
1	1	1	1	2	1	3	3

Child food source 2

1	2	3	4	5	6	7	8
1	1	2	2	2	2	3	3

Then choose the best solution with highest modularity fitness value in two child food sources. Equation 1 is used to

find better community quality. Child food source 2 get more fitness value than child food source 1. It is chosen as candidate solution. After getting candidate solution, choose new solution between original and candidate solution using greedy selection. To calculate mutation, each gene, pick up random probabilities, if the random probability less than mutation probabilities then goes on mutation. For genes to be mutated, replace the community index with neighbor node community index. After generating new food sources in employed bees' phases, calculate the probability of each food source and send this information to onlooker bees. Onlooker bee choose food source that based on probability of food sources. Then they also generate new solution as employed bee. If the solution is larger than maximum limit, scout bee searches new random solution as initialization stage. These process is iteratively done until the maximum cycle number. Finally, the algorithm produces the cluster member of nodes in the graph as an output. The sample graph output is community label 1(1,2,3), community label 2 (4,5,6) and community label 3(7,8) with best fitness 0.40123.

V. Experimental Results on Real Datasets

Famous datasets for community detection such as Zachary's karate club [10], Bottlenose dolphins [11], American college football [12] and US politics Book [13] networks are applied to show the effectiveness of proposed algorithm. For the algorithm, the parameter is set as number of population, p=100, maximum cycle number=200, limit=20, crossover rate =0.4 and mutation rate=0.1.

Traditional community detection algorithms are used to test on four datasets. This paper uses five traditional community algorithms infomap, label propagation, Leading Eigen vector, Louvain and walktrap [14-18] which are implemented in igraph library. That can be tested on RStudio by importing igraph package. Some algorithm cannot detect community structure in large scale graph. Nowadays, researchers use population based nature-inspired algorithms to find communities in large graph. Some nature-inspired algorithms are chosen to compare the quality of experimental results such as hybrid algorithm Teacher learners and group search optimization (TL-GSO), Group Search optimization (GSO-1), CNM, Evolutionary Algorithm (EA) Memetic algorithm [19-22].

Table 1 shows the comparison of modularity results get from various community detection algorithms. It shows that the proposed algorithm can detect suitable community structures in real graphs. In this table CNo is community number, M is modularity value of each algorithm. The result implies the best modularity value which leads to more accurate communities. Real world social graphs don't have ground truth community number. Researcher proposed algorithm to get effective results in choose objective function. Many of the proposed algorithms for community detection are based on modularity as an objective function.

TABLE I. Comparison of Modularity Values in Traditional Community Detection Algorithms and Nature-Inspired Based Algorithms

Algorithm	Zachary		Dolphin		Football		Book	
	CNo	M	CNo	M	CNo	M	Cno	M
Infomap	3	0.402	5	0.527	12	0.601	8	0.496
Label propagation	3	0.402	4	0.51	9	0.601	4	0.498
Leading eigenvector	4	0.393	5	0.491	8	0.492	4	0.465
Louvain	4	0.419	6	0.519	10	0.604	5	0.515
Walktrap	5	0.353	4	0.488	10	0.602	5	0.501
Proposed method	**4**	**0.419**	**4**	**0.526**	**9**	**0.603**	**4**	**0.52**
EA	3	0.38	3	0.46	7	0.56	4	0.52
TL-GSO	4	0.418	5	0.528	10	0.604	5	0.527
GSO-1	3	0.384	5	0.429	7	0.428	6	0.444
CNM	3	0.38	4	0.495	6	0.55	4	0.501
Memetic	3	0.402	4	0.518	7	0.604	4	0.523

According to the comparative modularity result, nature-inspired algorithms can be effectively solved community detection problem as traditional algorithms with suitable community number. They are also suitable for large-scale graphs. Spark framework is used to implement algorithm because it can be suitable for iterative machine learning task.

VI. Conclusion

In this paper, modified artificial bee colony with crossover and mutation is used to detected community in social graphs based on apache Spark. ABC is widely used in numerical optimization problem. It can also use in discrete problem as community detection. GraphX is used to create graph instead of adjacency matrix. GraphX is more suitable for large-scale graph processing. The proposed algorithm used modularity as a fitness function. In this paper, some conventional community detection algorithms and nature-inspired algorithms are applied to compare experimental results. The experiment results show that the proposed algorithm can effectively detect communities with quantitative modularity results and suitable community structures as other algorithms. This paper only considers edge structure of undirected graph. Future work will be considered to detect community on large-scale social graphs and GraphX on Apache Spark is also used for graph parallel processing.

ACKNOWLEDGMENT

I would like to think my supervisor, Dr. Thi Thi Soe Nyunt for her suggestion and helpful in my research and Dr. Khine Khine Oo, for her kindness and encourage. I also wish to thank all of my friends and family for their support, kindness and helpful.

REFERENCES

[1] Mustafa Servet Kiran, Hazim Iscan, Mesut Gunduz, " The analysis of discrete artificial bee colony algorithm with neighborhood operator on traveling saleman problem" , Neural & Applic(2013), Spinger

[2] https://igraph.org/r/doc/communities.html

[3] Shanfeng Wang, Maoguo Gong, YueWu and Xiaolei Qin, "Parallel Particle Swarm Optimization for Community Detection in Large-Scale Networks", Springer International Publishing AG 2017

[4] Gema Bello-Orgaz, David Camacho, "Evolutionary clustering algorithm for community detection using graph-based information", 2014 IEEE Congress on Evolutionary Computation (CEC) July 6-11, 2014, Beijing, China

[5] Ramadan Babers, Neveen I. Ghali, AboulElla Hassanein and Naglaa M. Madbouly, "Community Detection Based on Lion Optimization Algorithm", JOURNAL OF LATEX CLASS FILES, VOL. 6, NO. 1, JANUARY 2007

[6] YoucefBelkhiri, NadjetKamel, HabibaDrias, and SofianeYahiaoui, "Bee Swarm Optimization for Community Detection in Complex Network", Spinger International Publishing AG 2017.

[7] Ronghua Shang, Jing Bai, Licheng Jiao, Chao Jinm, " Communtiy detection based on modularity and an improved genetic algorithm", Physica A 392(2013) 1215-1231, 2012 Elsevier B.V.

[8] Rafael Stubs Parpinelli, C´esar Manuel Vargas Benitez, and HeitorSilv´erio Lopes, "Parallel Approaches for the Artifi cial Bee Colony Algorithm", DOI:10.1007/978-3-642-17390-5_14.

[9] SumanSaha and Satya P. Ghrera, "Network Community Detection on Metric Space", Algorithms 2015, 8, 1-13; doi:10.3390, www.mdpi.com/journal/algorithms

[10] W. Zachary. "An information flow model for conflict and fission in small groups." Journal of Anthropological Research, 33:452{473, 1977.

[11] D. Lusseau. "The emergent properties of dolphin social network." Proceedings of the Royal Society of London. Series B: Biological Sciences, 270: S186{S188, 2003.

[12] Girvan,M., Newman, M.E.J, " Community structure in social and biological networks.", Proc. Nat. Acad. Sci. 99(12), 7821-7826 (2002)

[13] V. Krebs, "A Network of Co-purchased Books About Us Politics", http://orgnet.com//

[14] M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences of the United States of America, 105(4):1118–23, Jan. 2008.

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

[16] Newman, M. E. "Finding community structure in networks using the eigenvectors of matrices". Physical Review E 74, 036104 (2006)

[17] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: "Fast unfolding of communities in large networks". J. Stat. Mech. (2008) P10008

[18] Pons, P. & Latapy, M. "Computing communities in large networks using random walks". In Computer and Information Sciences-ISCIS 2005, 284–293 (Springer, 2005).

[19] Banati, H., Arora, "N.: TL-GSO - a hybrid approach to mine communities from social network". In: 2015 IEEE International Conference

[20] Wu, P., Pan, L.: "Multi-objective community detection based on memetic algorithm." PLoS One 10(5), e0126845. doi:10.1371/journal. pone.0126845 (2015)

[21] Saoud Bilal, "Evolutionary Algorithm and modularity for Detecting Communities in Networks".

[22] Sunita Chand and Shikha Mehta, "Community Detection Using Nature Inspired Algorithm", © Springer International Publishing AG 2017, H. Banati et al. (eds.), Hybrid Intelligence for Social Networks