Information Sciences 316 (2015) 503–516

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Greedy discrete particle swarm optimization for large-scale social network clustering

Qing Cai, Maoguo Gong *, Lijia Ma, Shasha Ruan, Fuyan Yuan, Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 12 September 2014
Accepted 25 September 2014
Available online 7 October 2014

Keywords:
Social network
Clustering
Community structure
Particle swarm optimization
Large-scale global optimization

ABSTRACT

Social computing is a new paradigm for information and communication technology. Social network analysis is one of the theoretical underpinnings of social computing. Community structure detection is believed to be an effective tool for social network analysis. Uncovering community structures in social networks can be regarded as clustering optimization problems. Because social networks are characterized by dynamics and huge volumes of data, conventional nature-inspired optimization algorithms encounter serious challenges when applied to solving large-scale social network clustering optimization problems. In this study, we put forward a novel particle swarm optimization algorithm to reveal community structures in social networks. The particle statuses are redefined under a discrete scenario. The status updating rules are reconsidered based on the network topology. A greedy strategy is designed to drive particles to a promising region. To this end, a greedy discrete particle swarm optimization framework for large-scale social network clustering is suggested. To determine the performance of the algorithm, extensive experiments on both synthetic and real-world social networks are carried out. We also compare the proposed algorithm with several state-of-the-art network community clustering methods. The experimental results demonstrate that the proposed method is effective and promising for social network clustering.

1. Introduction

With the rapid development of the Internet and the Web 2.0 technologies, diverse types of social media networks, such as Blog, Wiki, Facebook, Twitter, and Weixin, have emerged and are changing the fundamental methods through which people share information and communicate. It is universally acknowledged that social media are profoundly affecting not only the global economy but also every aspect of our daily lives. Social media data are characterized by large data volumes, dynamics, interactivity and heterogeneity, which makes conventional computing models difficult to utilize. Social media data analysis challenges computing techniques in an unprecedented manner.

In 2009, scholars from the fields of social science, physical science, information science, etc., cooperated in an iconic work in [22], and since then, a new academic term has come into existence: social computing. Social computing, a new computing paradigm, is a cross-disciplinary research area integrating the computational and social sciences. According to Wikipedia,

* Corresponding author.
E-mail address: gong@ieee.org (M. Gong).

http://dx.doi.org/10.1016/j.ins.2014.09.041

social computing refers to the use of social software, the result of a growing trend in information and communication technology usage of tools that support social interaction and communication. Another definition can be found in [46]. Social computing aims at making use of computing techniques to help people to communicate and cooperate and to help to understand how a society operates and functions so as to direct decision making. The primary task for social computing is the model construction. Currently, the mainstream model is the social network model in which nodes denote the social objects, and edges represent social interactions among them. From this point of view, social computing largely depends on the analysis of the constructed social networks. Social networks have numerous features, among which the community structure is an eminent one. In academic domains, communities, also called clusters or modules, are groups of vertices that most likely share common properties and/or play similar roles within the graph. Probing the community structures of networks can help us to understand how a network functions.

The discovery of community structures in networks can be considered as an optimization problem [29]. In the past few decades, numerous nature-inspired optimization algorithms characterized by good local learning and by global searching abilities have gathered momentum through both theoretical and empirical studies. Evolutionary algorithms (EAs) are some representative optimization paradigms among the nature-inspired optimization algorithms. EAs have been developed and successfully applied to a wide range of optimization problems, including network clustering optimization. Recently, scholars have successfully applied both single- and multi-objective EAs to discover community structures in networks [32,33,13]. In addition to the EA-based metaheuristic optimization techniques, another class is swarm-intelligence-based avenues, among which the outstanding paradigm is particle swarm optimization (PSO) [18]. PSO originated from the behavior of social animals, such as fish schooling and birds flocking. PSO optimizes a problem by employing a group of particles. Each particle is a candidate solution to the problem. The candidate solutions are updated with simple rules learnt by the particles. Due to its efficacy and its extremely easy implementation, PSO is prevalent in the optimization field, and diverse variants have been proposed [43,28,42]. However, canonical PSO is specially designed for continuous optimization problems. Although some discrete successors have emerged in the literature, their applications are still limited.

Note that due to the huge volume of data sets, social network clustering optimization is a large-scale global optimization (LSGO) problem. LSGO problems refer to optimization problems that involve a large number of decision variables. LSGO is difficult for existing optimization techniques. Studies on scaling up nature-inspired optimization algorithms to solve LSGO problems have attracted much attention [53,47,50,23]. In this paper, a discrete PSO algorithm for large scale social network clustering is put forward for the first time. The main highlights of the proposed algorithm are as follows. First, in order to handle the LSGO network clustering problem, the particles' velocity and position have been carefully redefined under discrete context so as to make them as easier as possible to encode/decode. Second, for the sake of better exhaustive global searching in the vast searching space, to drive the particles to promising regions, the particle-status-update principles have been thoroughly reconsidered by taking the advantage of the network topology. Third, to avoid being trapped into local optima, a greedy strategy specially designed for the particles to adjust their positions is newly suggested. The proposed algorithm is denoted by GDPSO. To check the performance of GDPSO, experiments on computer-generated and real-world social networks are done. On the real-world networks, we test GDPSO on both small scale and large scale networks. We also compare GDPSO with several state-of-the-art network clustering methods. The experimental results show that the proposed GDPSO algorithm is very effective and promising.

The remainder of this paper is organized as follows. Section 2 illustrates the related background and gives the motivation for this work. In Section 3, the proposed GDPSO framework is presented in detail. The designs of the particle representation and the update principles are described. Section 4 shows the performance testing of the proposed method, and the concluding remarks are subsequently summarized in Section 5.

2. Related background

2.1. Network clustering

Social networks are constructed based on social behaviors and relations. In graph theory, a social network can be expressed by a graph that is composed of nodes and edges, as shown in Fig. 1a. Social network clustering separates the whole network into small parts, within which the similarities are maximized and between which the dissimilarities are maximized, as in the toy model shown in Fig. 1b. These small parts are called communities in the literature.

Based on the node degree, Radicchi et al. in [34] gave a qualitative definition of a community. Let $G = (V, E)$ denote a network, where V and E are the aggregations of vertices and links, respectively. Let $k_i^{in} = \sum_{i,j \in S} A_{ij}$ and $k_i^{out} = \sum_{i \in S, j \notin S} A_{ij}$ be the internal and external degree of node i, A is the adjacency matrix of G, and $S \subset G$ is the subgraph that node i belongs to. Then, S is a community in a strong sense if $\forall i \in S,\ k_i^{in} > k_i^{out}$ and in a weak sense if $\sum_{i \in S} k_i^{in} > \sum_{i \in S} k_i^{out}$.

Thus far, a large number of clustering methods have been developed to discover communities in networks. These clustering methods can be categorized into three classes: (1) hierarchical clustering methods [20,15], which iteratively merge or split clusters according to vertex similarities; (2) partitional clustering approaches, such as the k-means [8] and fuzzy k-means methods [52]; and (3) spectral clustering methods [30,49], which use the spectra information of the adjacency matrix to partition networks. A recent survey on network community detection can be found in [9].