Information Sciences 267 (2014) 101–118

Contents lists available at ScienceDirect

Information Sciences

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

# A global optimization algorithm for target set selection problems

Cheng Wang [a], Lili Deng [b,*], Gengui Zhou [b], Meixian Jiang [a]

[a] Institute of Industrial Engineering, Zhejiang University of Technology, Hangzhou 310014, China
[b] College of Economics and Management, Zhejiang University of Technology, Hangzhou 310023, China

ARTICLE INFO

*Article history:*
Received 29 June 2012
Received in revised form 13 May 2013
Accepted 13 September 2013
Available online 21 September 2013

*Keywords:*
Target set selection
Genetic algorithm
Diffusion of information
Marketing
Social network

ABSTRACT

Our study concerns the target set selection problem, which involves discovering a subset of influential players in a given social network performing a task of information diffusion to maximize the number of nodes influenced in the network. We are motivated by the facts that the well-known algorithms for target set selection problems are heuristic, and the best heuristic algorithm only ensures that the spread is within 63% of the optimal influence spread based on the submodular assumption. We propose a set-based coding genetic algorithm (SGA), which converges in probability to the optimal solution of target set selection problems. Computational experiments on four synthetically generated graphs and five real-world data sets are carried out to compare the performance of the proposed SGA with those well-known algorithms in the literature. Statistical significance tests indicate that the proposed SGA outperforms the state-of-the-art algorithms for target set selection problems significantly.



## 1. Introduction

The process by which new ideas or behaviors spread through a population has long been a fundamental question in social sciences [21], such as new religious beliefs, the adoption of new technologies, and the success of a new product. These phenomena all share some properties: They tend to begin on a small scale with a few early adopters; then more and more people begin to adopt them as they observe their friends, neighbors, or colleagues doing so; and the resulting new behaviors may eventually spread through the population [21]. Such processes are studied as the diffusion of information. Schelling [30] and Granovetter [16] formulate the basic mathematical models for the mechanisms by which new ideas and behaviors diffuse through a population.

During the diffusion process, the social network plays a fundamental role as a medium [8]. In a social network, the underlying representation is a graph model where the nodes represent individuals and the edges represent interactions between them. Studies of the structure and dynamics of social networks are still in their infancy [6,33,36]. In the 1980s and 1990s, researchers, such as Axelrod [2] and Cowan and Miller [9], favor simple networks coupled in geometrically regular ways. However, this regular network has limitations for research that attempts to capture essential aspects of complex social networks. Another simple network topology is random network [15], in which an individual can be randomly connected to any other individual in the world. However, studies indicate that many real-world networks systematically deviate from the topology predicted by the random network theory [27]. Watts and Strogatz [35] offer a clue to tackle the complexity inherent in the structures of real networks by the idea that a complex network in the real world may lie between a regular network and a random network. Watts and Strogatz [35] formalize this idea by developing an algorithm that can encompass the span of possible topologies between a regular network and a random network.

* Corresponding author.
*E-mail addresses:* cwang@zjut.edu.cn (C. Wang), denglili@zjut.edu.cn (L. Deng), ggzhou@zjut.edu.cn (G. Zhou), jmx@zjut.edu.cn (M. Jiang).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.09.033

Two popular operational models, the linear threshold model and the independent cascade model, capture the dynamics of the diffusion process in a social network in the literature.

The linear threshold model is proposed by Granovetter [16] and generalized by Watts [34]. In this model, each node $i$ of the network is associated with one of the two states, active or inactive, which indicates whether individual $i$ adopts the information (new ideas or behaviors) or not. A nonnegative weight $w_{ij}$ is assigned to each edge $(i,j)$, indicating the influence that $j$ exerts on $i$. It is required that $\sum_{j\in N(i)} w_{ij} \leqslant 1$, where $N(i)$ denotes the set of nodes, which connect to $i$. Moreover, individual $i$ is assigned a threshold value $\theta_i$, chosen randomly from the interval $[0,1]$. The threshold value specifies the weighted fraction of $i$'s neighbors that must adopt the information before $i$ does. For instance, if $A(i)$ is assumed to be the set of active neighbors of node $i$, then $i$ will get active if $\sum_{j\in A(i)} w_{ij} \geqslant \theta_i$. The diffusion process runs in the network as follows: Initially, all nodes are inactive. In phase 0 of the process, a set of nodes are set to be active. At phase $t > 0$, a node $i$ becomes activated if the total weight of its active neighbors is at least $\theta_i$. Once a node becomes activated, it remains active for the entire process. This process continuous until no new nodes get activated.

In the independent cascade model, the diffusion process works as follows: A set of nodes $A(0)$ are initialized to be active. Let $A(t)$ be the set of nodes that are activated at the $t$th round. For any edge $(i,j)$, where $i \in A(t)$ and $j$ is inactive, $j$ gets activated by $i$ in the $(t+1)$th round with a probability independent of the history thus far. If $j$ has multiple newly activated neighbors, their attempts are sequenced in an arbitrary order. This process is repeated until $A(t+1)$ is empty.

Given a social network and a diffusion dynamic, a natural question arises: how to find a target set of desired cardinality that consists of influential nodes for maximizing the volume of the information cascade [11,19,20]. We call such a set of influential nodes as a target set, and the problem of determining a target set of particular cardinality to perform a given task in the social network as a target set selection problem [5]. One example is viral marketing, where a company tries to use word-of-mouth effects to market a product with a limited advertising budget, relying on the fact that early adopters may convince their friends and colleagues to use the product, creating a large wave of adoptions.

Concretely, the target set selection problem can be formulated as follows. For any given diffusion dynamic, there is a natural influence function $f(\cdot)$ such that for a set of active nodes $S$, $f(S)$ is the expected number of activated nodes at the end of the process. From the marketer's point of view, $f(S)$ is the expected number of total sales if $S$ is the set of initial adopters. Now, how large $f(S)$ can be if we are allowed to choose a set $S$ of $k$ initial adopters? In other words, we try to maximize $f(S)$ over all subsets of size $k$.

### 1.1. Related works

Domingos and Richardson [11] and Richardson and Domingos [28] are the prime works to study the target set selection problem, where social networks are modeled as Markov random fields. Kempe et al. [19,20] are the first to formulate the problem as a discrete optimization problem. Three cascade models, namely the independent cascade model, the weight cascade model, and the linear threshold model, are considered by Kempe et al. [19]. They prove that the optimization problem is NP-hard. Considering the resulting influence function $f(\cdot)$, they first show that this function is submodular under the independent cascade model and the linear threshold model. Formally, $f(\cdot)$ is said to be submodular if it satisfies $f(S\cup\{i\}) - f(S) \geqslant f(T\cup\{i\}) - f(T)$, for all elements $i$ and all pairs of sets $S \subseteq T$. Kempe et al. [19] propose a greedy hill-climbing algorithm described in Algorithm 1.1, which guarantees that the influence spread is within $(1-1/e) \approx 63\%$ of the optimal influence spread based on the submodular assumption.

**Algorithm 1.1.** Greedy Algorithm of Kempe et al. [19]. $N$ is the set of nodes and $k$ is a positive integer such that $k \leqslant |N|$

1: $A \leftarrow \phi$;
2: **for** $i = 1$ to $k$ **do**
3: Choose a node $n_i \in N\backslash A$ maximizing $f(A\cup\{n_i\}) - f(A)$;
4: $A \leftarrow A \cup \{n_i\}$;
5: **end for**

Leskovec et al. [22] propose the Cost-effective lazy forward selection (CELF) optimization to the original greedy algorithm based on the submodularity of the influence maximization objective. The CELF optimization has the same influence spread as the original greedy algorithm but is 700 times faster than the greedy algorithm as reported by Leskovec et al. [22].

Chen et al. [7] present an efficient algorithm, which improves the greedy algorithm of Kempe et al. [19] and also the CELF optimization of Leskovec et al. [22] in terms of its running time. Besides, Chen et al. [7] also design a degree discount heuristic algorithm, which achieves much better influence spread than classic degree and centrality-based heuristics. Meanwhile, the performance of degree discount heuristics is comparable to that of the greedy algorithm while its running time is much less than that of the greedy algorithm.

Inspired by cooperative game theory, Narayanam and Narahari [26] propose the Shapley value-based influential nodes (SPIN) algorithm for solving the diffusion maximization problem. This algorithm mainly includes two steps: compacting a ranking list of the nodes based on Shapley value, and choosing the top $k$ nodes from the rank list. They compare the performance of the SPIN algorithm with the greedy algorithm when the influence functions are submodular as well as when the