ARTICLE IN PRESS

Swarm and Evolutionary Computation xxx (2017) 1–8

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

Finding influential users for different time bounds in social networks using multi-objective optimization

Azadeh Mohammadi [a,*], Mohamad Saraee [b]

[a] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran
[b] School of Computing, Science and Engineering, University of Salford-Manchester, Manchester, United Kingdom

ARTICLE INFO

Keywords:
Diffusion time
Influence maximization
Multi-objective optimization
Social networks
Trade-off decision making

ABSTRACT

Online social networks play an important role in marketing services. Influence maximization is a major challenge, in which the goal is to find the most influential users in a social network. Increasing the number of influenced users at the end of a diffusion process while decreasing the time of diffusion are two main objectives of the influence maximization problem. The goal of this paper is to find multiple sets of influential users such that each of them is the best set to spread influence for a specific time bound. Considering two conflicting objectives, increasing influence and decreasing diffusion time, we employ the NSGA-II algorithm which is a powerful algorithm in multi-objective optimization to find different seed sets with high influence at different diffusion times. Since social networks are large, computing influence and diffusion time of all chromosomes in each iteration will be challenging and computationally expensive. Therefore, we propose two methods which can estimate the expected influence and diffusion time of a seed set in an efficient manner. Providing the set of all potentially optimal solutions helps a decision maker evaluate the trade-offs between the two objectives, i.e., the number of influenced users and diffusion time. In addition, we develop an approach for selecting seed sets, which have optimal influence for specific time bounds, from the resulting Pareto front of the NSGA-II. Finally, we show that applying our algorithm to real social networks outperforms existing algorithms for the influence maximization problem. The results show a good compromise between the two objectives and the final seed sets result in high influence for different time bounds.

1. Introduction

Millions of people around the world subscribe to popular social networks such as Facebook, twitter, google+ and other social media. The high usage of social networks provides a good opportunity for social network analysis. There are many different research directions in social network analysis, including information diffusion [1–3], community detection [4–7], link prediction [8–10] and behaviour analysis [11,12].

One of the main research interests in social networks is analysing information diffusion which is an important mechanism for viral marketing. Viral marketing is a marketing strategy that is based on the influence between individuals such as families, friends or colleagues [13,14]. Sociological studies have shown that information coming from close relatives has a strong impact on an individual's decision and in this way a trend, behaviour or information can propagate on the network [1]. Consequently, finding the most influential users in a social network becomes a critical task. The challenge of finding the k most influential users in a social network is called influence maximization [3]. The set containing these influential users is called a seed set. The activation of this set of nodes can maximize the expected spread of information in the network.

Influence maximization is widely studied and many different approaches have been considered to identify scalable solutions for finding the most influential users [15–19]. Many of these approaches do not consider the time required to achieve influence spread during the maximization process. But, in some cases, time could be a criterion and has various levels of importance in different marketing applications. For instance, marketers in a clothing company have to advertise and maximize the number of sales of the winter products in cold seasons. As an example, customers may become aware of winter collection of boots and coats from early winter. There is a limited time for advertisement of such products. In this case, in addition to influence spread, i.e., the number of active users, diffusion time also becomes critical. In other applications, time may be less important, therefore, maximizing the influence over

* Corresponding author.
E-mail address: amohammadi@ec.iut.ac.ir (A. Mohammadi).

https://doi.org/10.1016/j.swevo.2018.02.003
Received 8 June 2017; Received in revised form 27 December 2017; Accepted 7 February 2018
Available online xxxx

unlimited time would suffice.

Diffusion time is the elapsed time from the beginning of information spread until no other user can be activated. This time depends on the delay (time taken) by each user to spread information. Since users have different propagation delays, different seed sets may reach their final influence spread after spending different diffusion times. Liu et al. [17] showed that a seed set which can maximize the amount of influence without considering any time bound is not necessarily the best choice for those applications which have a specific time bound.

Given that increasing influence and decreasing diffusion time are two conflicting objectives, there would be several seed sets each having the best influence for a specific time bound. Finding all these sets at the same time will help decision makers to select the best seed set that fits their needs.

The purpose of this paper is to find near-optimal seed sets for different time bounds simultaneously. We have sought solutions that have more influence in specific time bounds and have compared our results with other approaches.

The major contributions of our paper are as follows:

- We show the relationship between influence and time in the information diffusion process and consider a novel problem which aims to find near-optimal seed sets for different time bounds simultaneously.
- We propose an algorithm which is able to select the optimal seed sets from the results of NSGA-II algorithm [20].
- We propose two heuristics to estimate the amount of influence spread and diffusion time in the NSGA-II algorithm in order to reduce the running time of the algorithm.
- We compare the results of our approach with an existing algorithm for influence maximization which considers time limitation in its calculations and also with two other heuristic methods, on three real datasets.

This paper is organized as follows. In section 2, we briefly review research in the area of the influence maximization problem and explain multi-objective optimization and the NSGA-II algorithm. In section 3, we present our proposed method and explain how we have employed a multi-objective solution for our problem and how we have handled the large scale issue of social networks. In section 4, we show the results of applying our approach on real datasets. Conclusion of our work and future work are discussed in section 5.

2. Related works

The influence maximization problem aims to find k nodes in a social network that directly and indirectly influence the largest number of nodes under a predefined diffusion model. Richardson et al. [21,22] studied the problem of influence maximization and modelled it using the Markov random field method. This problem was first formulated as a discrete optimization problem by Kempe et al. in 2003 [16]. They modelled a social network as a directed graph, in which nodes are individuals and edges represent relationships between them. In their work, they presented two stochastic diffusion models: an independent cascade (IC) model and a linear threshold (LT) model. These have become the baselines for other works on the influence maximization problem. In addition, they proved that the influence maximization problem is NP-hard and proposed a greedy algorithm with a (1-1/e) approximation guarantee of an optimal solution.

It is proved that the problem of calculating the exact influence spread of a seed set is NP-hard under IC [23] and LT [24] models. Due to the complexity of the simulation approach to find influence spread, a set of algorithms such as CELF [25], CELF++ [26], STATICGREEDY [27], RIS [28] and TIM [29] are proposed that improve the runtime of Kempe's algorithm, while maintaining the influence spread guarantee.

In contrast to approximation algorithms, several studies use heuristic methods to select influential nodes in the influence maximization problem, among which, we can mention LDAG [24] and SIMPATH [30] for the LT model and MIA [23,31] and IRIE [32] for the IC model.

Most of the previous work in this area has focused only on the coverage of seed set. Tang et al. [33] introduced node diversity as a second objective function. They defined diversified social influence maximization, formulated it as an optimization problem and provided a greedy solution. Gunasekara et al. [20] used the NSGA-II algorithm to identify the set of key players in the social networks by considering two objectives, eigenvector centrality and the distance between key players.

Very recently, some works have considered the time aspect in the diffusion process. Gomez-rodriguez et al. [34,35] proposed methods for learning the network structure and transmission rates using temporal propagation data. Goyal et al. [36], suggested methods for estimating the parameters of static and time-dependent models from the propagation data. Mohammadi et al. [37], introduced the Time-Sensitive Influence Maximization problem, in which the dependency of information value on time is incorporated into the influence maximization process. They noted that in time-sensitive applications, not only the coverage size, but also the time of activation matters. Liu et al. [17] introduced time constrained influence maximization in which the goal is to find a set of users which can maximize the amount of influence in a limited time bound. They defined an influence spreading path concept and proposed scalable algorithms that can be scaled to large social networks. However in their work, one needs to run such an algorithm with different time bounds in order to find suitable seed sets for each particular time bound. The selection of seed nodes in the Liu's approach is done in a greedy manner. A node will be added to the current seed set if it gives the highest influence gain in a limited time bound. This process is repeated until the desired number of seeds is achieved.

In our work, we have considered both the coverage size and the diffusion time as the objectives of influence propagation. In fact, increasing coverage size and decreasing diffusion time are two conflicting objectives that we consider in the selection of seed nodes. This introduces a new multi-objective optimization problem for which we employ NSGA-II algorithm as a solution. Given the massive scale of social networks, we propose two efficient methods for estimating influence and diffusion time to reduce the execution time of NSGA-II in our problem.

Furthermore, we discuss the results under time constraints and propose an algorithm for selecting influential nodes in specific time bounds. We compare the influence spread of our approach with results obtained by Liu et al. [17], for specific time bounds. The details of experiments are given in section 4.

2.1. Influence maximization and diffusion models

In this section, we present the influence maximization problem and the Independent Cascade (IC) model. In addition, we discuss the time constrained influence maximization problem and explain a latency-aware IC model, which considers time in the diffusion process.

Definition 1. Given a directed social network $G = (V, E)$, where V represents the set of nodes and E represents the set of edges between them, positive number $k < |V|$ and propagation probability P_{uv} between each edge uv, the goal of the influence maximization problem is to find the seed set S containing k nodes where the expected number of active nodes using S, $\sigma_T(S)$, is maximized at the end of the diffusion process.

Several models have been presented to demonstrate how influence propagates in a social network [16,38,39]. One of the most popular models for describing influence propagation is the independent cascade (IC) model which is presented by Kempe et al. [16]. In this model, each node can be in active or inactive state. Inactive nodes can become active but active nodes will stay active. In a diffusion process based on this model, each active node u which is activated at step t, has a single chance to activate all of its currently inactive neighbours at step $t+1$. An active node u can activate its inactive neighbour v with probability $P_{uv} \in [0, 1]$ which is called propagation probability and is assigned on the edge that