Information Sciences 316 (2015) 107–122

Contents lists available at ScienceDirect

Information Sciences

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

Deprecation based greedy strategy for target set selection in large scale social networks

Suman Kundu *, Sankar K. Pal

Center for Soft Computing Research, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 5 July 2014
Received in revised form 2 April 2015
Accepted 11 April 2015
Available online 17 April 2015

Keywords:
Top-*k* nodes selection
Social network
Greedy deprecation strategy
Influence maximization
Big data
Target set selection

ABSTRACT

The problem of target set selection for large scale social networks is addressed in the paper. We describe a novel deprecation based greedy strategy to be applied over a pre-ordered (as obtained with any heuristic influence function) set of nodes. The proposed algorithm runs in iteration and has two stages, (i) Estimation: where the performance of each node is evaluated and (ii) Marking: where the nodes to be deprecated in later iterations are marked. We have theoretically proved that for any monotonic and sub-modular influence function, the algorithm correctly identifies the nodes to be deprecated. For any finite set of input nodes it is shown that the algorithm can meet the ending criteria. The worst case performance of the algorithm, both in terms of time and performance, is also analyzed. Experimental results on seven un-weighted as well as weighted social networks show that the proposed strategy improves the ranking of the input seeds in terms of the total number of nodes influenced.

1. Introduction

Information diffusion over social networks in the form of "word-of-mouth" is studied in different fields of research including epidemiology [9,34], sociology [2,39] and economics [15,16]. More recently, scholars of computer science got interested in the field due to the emergence of online social networks, like, Twitter, Facebook and YouTube, and their extreme popularity. Different research issues have been addressed in this direction [11,18,20,21,38,41]. One of the important problems within the area of said research is *target set selection*.

A variant of the problem of target set selection is to select *k*-top influential nodes such that they maximize the influence on the network. There are other variants in the literature such as those in [3,29] which we will not cover in this study. Solutions of the target set selection problem have endless applications. For example, they are useful in viral marketing through online social networks [10,27], in identifying top stories in news network, in finding the highest influencing blogs in the blogger network [26], in providing personalized recommendation [17,40], in determining the impact of an article from the scientists' citation network, and in spreading social awareness through social media.

Diffusion of information, in a nutshell, is the process by which an innovation or idea is spread over the networks by means of communication among the social entities [36]. It is the newness of the information that drives the cascade over the networks. One of the simplest models of the diffusion process available for the computer science researchers is *independent cascade model* [15,20]. The model runs in discrete steps. In each step, an active or influenced node tries to activate one of its

* Corresponding author.
E-mail addresses: suman@sumankundu.info (S. Kundu), sankar@isical.ac.in (S.K. Pal).

http://dx.doi.org/10.1016/j.ins.2015.04.024
0020-0255/

inactive neighbors with a probability p, called propagation probability or diffusion speed. Irrespective of its success, the same node will never get a chance to activate the same neighbor. The process, however, is highly stochastic, and Kempe et al. [20] showed that the optimization problem is NP hard. They also provided a Greedy Hill Climbing algorithm, which gives $(1-\frac{1}{e}-\epsilon)$ approximation to the optimal solution. However, the algorithm is time consuming, especially for large scale networks. For example, it takes days to compute on a network of size 30 K nodes [6]. Various improvements of the greedy algorithm in terms of computation time are described in [3,12,24]. On the other hand, several heuristic algorithms [3,4,6] are developed which run faster, but they provide sub-optimal results.

This paper addresses the aforesaid problem within the context of information diffusion on large scale social networks. We describe a *deprecation based greedy strategy* (DGS) for target set selection and apply it over a list of nodes which are pre-ordered based on some fast heuristic influence score. We theoretically prove that the method correctly identifies the nodes to be deprecated as well as provides a guaranteed solution to the target set selection problem when the influence function is monotonic and sub-modular. The convergence of the proposed algorithm is proved analytically. It is shown experimentally, with seven real life large scale social network data sets (both weighted and un-weighted) that applying DGS over a heuristic algorithm produces better solution for the target set selection problem.

The paper is organized as follows: Section 2 describes the preliminary concepts of networks related to this study. Problem statement and related investigations are briefly explained in Section 3. Section 4 illustrates the Deprecation based Greedy Strategy (DGS), and its proof of correctness, convergence and optimization guarantee. Experiments and results are reported in Section 5.

2. Preliminaries

We describe in this section some notations and definitions related to social networks.

2.1. Social networks

A social network represents a social structure made up of individuals or organizations and their relations (e.g., friendship, co-authorship of scientific papers, co-appearance in a movie, and following-followers). Social networks are described using a graph $G(V,E)$ where V is the set of nodes representing the individuals or organizations and E is the set of edges representing the social ties.

2.2. Information diffusion

Information diffusion process and the effect of "word-of-mouth" in social networks is well studied in sociology [36]. During the diffusion process there exist two sets of nodes, namely, active and inactive nodes. The active nodes are those who have already adapted the behavior, i.e., have the information, while the inactive nodes are those who do not have. One of the fundamental processes of information diffusion available in the literature is cascade model. Goldenberg et al. [15,16] inspected cascade model in the marketing perspective. In this model, a node u is influenced by its neighbor v with a probability $\lambda_{u,v}$, called propagation probability.

2.2.1. Independent cascade (IC) model

The IC model of [15] is the simplest form of the cascade model of diffusion and runs in discrete time. Initially, a few nodes are activated. At each successive step, an active node tries to activate one of its inactive neighbors. The node, however, gets only one chance to activate that particular node irrespective of its success. The process terminates when no further activation is possible. Edge $e(u,v) \in E$ is assigned a non-negative propagation probability $\lambda_{u,v}$ which indicates the probability at which node u is activated by v.

2.3. Centrality

In a social network, centrality of a node provides a measure of its relative importance in the network. This is considered to be an important structural attribute [14] of the network. Two ways of measuring it are as follows:

- *Degree centrality:* Degree centrality of a node v is defined in terms of the numbers of incident edges as [30],

$$C_D(v) = \sum_{i=1}^{n} e(u_i, v) \tag{1}$$

where

$$e(u_i, v) = 1, \text{ if the nodes } u_i \text{ and } v \text{ are connected, i.e., an edge exists between them, and} = 0, \text{ otherwise.}$$

- *Diffusion degree centrality:* The diffusion degree of a node v is defined as [33],