Expert Systems with Applications 42 (2015) 1353–1361

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A fast algorithm for finding most influential people based on the linear threshold model

Khadije Rahimkhani [a,1], Abolfazl Aleahmad [a,1,*], Maseud Rahgozar [a], Ali Moeini [b]

[a] Database Research Group, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Iran
[b] Faculty of Engineering Science, School of Engineering, University of Tehran, Iran

ARTICLE INFO

Article history:
Available online 30 September 2014

Keywords:
Social networks
Influential people retrieval
Influence maximization
Linear threshold model

ABSTRACT

Finding the most influential people is an NP-hard problem that has attracted many researchers in the field of social networks. The problem is also known as influence maximization and aims to find a number of people that are able to maximize the spread of influence through a target social network. In this paper, a new algorithm based on the linear threshold model of influence maximization is proposed. The main benefit of the algorithm is that it reduces the number of investigated nodes without loss of quality to decrease its execution time. Our experimental results based on two well-known datasets show that the proposed algorithm is much faster and at the same time more efficient than the state of the art algorithms.

1. Introduction

Social networks play an important role in our new world. After invention of online social networks, people are able to influence each other much more easily. This fact caused many researchers to look for an efficient method for finding top-k most influential people through social networks. This problem is also known as influence maximization (IM) and has many applications such as: opinion propagation, studying acceptance of political movements or acceptance of technology in economics. For example, suppose that we need to advertise a product in a country or we need to propagate news. For this purpose, we need to choose some people as a starting point and maximize the news or the products influence in the target society.

Finding most influential people has been found useful for many applications such as: developing recommender systems (Morid, Shajari, & Hashemi, 2014; Song, Tseng, Lin, & Sun, 2006), choosing useful weblogs (Leskovec et al., 2007) and finding influential twitters (Bakshy, Hofman, Mason, & Watts, 2011; Weng, Lim, Jiang, & He, 2010). The problem was introduced in (Domingos & Richardson, 2001) for the first time. After that in (Kempe, Kleinberg, & Tardos, 2003) the authors formalized the problem as follows: given a weighted graph in which nodes are people and edge weights represent influence of the people on each other, it is desired to find K starting nodes that their activation leads to maximum propagation based on a chosen influence maximization model.

Inspired from humanities science, two well-known influence maximization models were presented for the first time in (Kempe et al., 2003): Linear Threshold (LT) and Independent Cascade (IC). Also, different algorithms have been proposed based on IC model (Morid et al., 2014; Wang, Cong, Song, & Xie, 2010; Zhang, Zhou, & Cheng, 2011) and LT model (Brin & Page, 1998; Chen et al. 2010; Goyal, Bonchi, & Lakshmanan, 2011; Leskovec et al., 2007). In this paper, a new algorithm will be introduced for finding top-k most influential people based on the LT model. Also, through experimental results on two real world datasets, we will show that the proposed algorithm is faster and even more efficient than the state of the art algorithms, so it would be applicable to larger social networks.

In this paper, we focus on selecting the most important nodes that lead to maximum influence based on the LT model. For this purpose, community structures of the input graph are identified first and the most influential communities are selected among them; then a number of representative nodes are chosen from the resulted communities and form the final output of the proposed algorithm. Our experiments show that the proposed algorithm is very efficient in finding the most influential nodes with minimum execution time. In other words, in this paper we will focus on optimizing both the influence spread and the execution time.

* Corresponding author.
E-mail addresses: rahimkhani_khadije@yahoo.com (K. Rahimkhani), a.aleahmad@ece.ut.ac.ir (A. Aleahmad), Rahgozar@ut.ac.ir (M. Rahgozar), moeini@ut.ac.ir (A. Moeini).
[1] The first two authors contributed equally to this work.

http://dx.doi.org/10.1016/j.eswa.2014.09.037
0957-4174/

The rest of the paper is organized as follows: the next section reviews the related works; the proposed algorithm is presented in Section 3; comparison and analysis of the results are presented in Section 4 and finally, Section 5 concludes the paper.

2. Related works

As stated in the previous section, two different information diffusion processes are presented in (Kempe et al., 2003): IC and LT models. In IC model, when a node is activated in time t, it has only one chance to activate its neighbors in time $t+1$. In LT model, in-links of each node are weighted with a value less than one and each person has an activation threshold between zero and one. Then, if total influence rate of neighboring nodes of a node is greater than its threshold, the node will be activated.

The influence function is monotone and sub-modular in both models. The function set f: $2^U \rightarrow R^+$ is monotone if $f(S) \leqslant f(T)$ for all $S \subseteq T \subseteq U$. Also, if $f(S \cup \{w\}) - f(S) \geqslant f(T \cup \{w\}) - f(T)$ is true for all $S \subseteq T$ and $w \in U \setminus T$, then the function is sub-modular. So, both IT and IC models of influence maximization are NP-hard problems. Furthermore, in (Lu, Zhang, Wu, Kim, & Fu, 2012) the authors study the complexity of the influence maximization problem in deterministic linear threshold model. They show that there is no $n^{1-\epsilon}$ – factor polynomial time approximation for the problem in the deterministic linear threshold model unless P = NP. Also, they show that the exact computation of the influence given a seed set can be solved in polynomial time.

The rest of this section reviews LT and IC based influence maximization algorithms. Then the most important community detection algorithms are discussed and finally a formalization of linear threshold model will be presented.

2.1. IC-based influence maximization algorithms

There exist different solutions for Independent Cascade (IC) model of the problem. In (Guo, Zhang, Zhou, Cao, & Guo, 2013), the authors present a dynamic selection approach referred to as the Item-based top-k influential user Discovering Approach (IDA). IDA identifies the top-k influential users for a given topic based on real-world diffusion traces and on-line relationships. In (Kim, Kim, & Yu, 2013), the authors propose a scalable influence approximation algorithm named Independent Path Algorithm (IPA). In (Li et al., 2014), the polarity-related influence maximization (PRIM) problem is proposed which aims to find the seed node set with maximum positive influence or maximum negative influence in signed social networks. In (Liu et al., 2013), the authors propose GS algorithm for quick approximation of influence spread by solving a linear system, based on the fact that propagation probabilities in real-world social networks are usually quite small. Furthermore, for better approximation, they study the structural defect problem existing in networks, and correspondingly, propose enhanced algorithms, GSbyStep and SSSbyStep, by incorporating the Maximum Influence Path heuristic. In (Yang et al., 2012), the authors propose to measure the seed's independent influence by a linear social influence model. In (Li et al., 2014), a novel conformity aware cascade model is proposed which leverages on the interplay between influence and conformity in obtaining the influence probabilities of nodes from underlying data for estimating influence spreads. Also, in (Ohsaka, Akiba, Yoshida, & Kawarabayashi, 2014), the authors propose a new method that produces solutions of high quality base on Monte Carlo simulation.

2.2. LT-based influence maximization algorithms

In (Leskovec et al., 2007), the authors introduced a sub-modularity factor and presented a lazy-forward greedy algorithm named CELF. Then Goyal et al. presented CELF++ that is an extension of CELF which reduces the number of iterations of the algorithm considerably (Goyal, Lu, & Lakshmanan, 2011a).

There exists different works on influence models such as: (Chen et al. 2010; Goyal, Lu, & Lakshmanan, 2011b; Kimura & Saito, 2006). Simpath algorithm (Goyal et al., 2011b) estimates activation probability of nodes by investigating paths that exists between seed nodes and other nodes in the input network. Kempe et al. used a greedy algorithm to add K nodes with maximum marginal gain to the seeds set (Kempe et al., 2003). Exact calculation of marginal nodes is #P-hard in both IC and LT models (Bakshy et al., 2011; Chen et al., 2010). Hence, they are estimated by use of Monte Carlo (MC) simulation. Unfortunately, the greedy algorithm has the following drawbacks:

- MC simulation should be run for many times (e.g. 10,000 times).
- This greedy algorithm should use MC for $n \times k$ times. Where n is the number of nodes and k is the number of selected nodes.

Chen et al. came to the conclusion that based on LT models, influence maximization problem is #P-hard (Chen et al., 2010) and it can be executed on directed acyclic graphs (DAGs) in linear time. They supposed that each node can only influence a limited number of its neighbors. So, a local DAG (LDAG) is considered for each node and its influence is investigated for its LDAG. They showed that this heuristic algorithm is faster and even more efficient than its greedy counterparts. However, their idea has the following limitations:

- Finding LDAGs is an NP-hard problem, so a greedy heuristic is employed to discover a good LDAG in (Chen et al., 2010). Using a greedy LDAG may introduce loss in quality of the seed set.
- It is supposed that a node can influence others through the paths that exist in its LDAG, so its influence on other paths is ignored.
- Storage of all LDAGs needs a huge memory in large networks.

This algorithm is very efficient for LT model of influence maximization. It is faster than greedy algorithms like CELF but it does not propose high quality starting nodes. Another algorithm is presented in (Leskovec et al., 2007) that considers another condition for simple greedy algorithms. They presented an optimization for CELF that reduces the number of MC iterations considerably. Despite this optimization, the greedy algorithm needs nearly 700 iterations (Leskovec et al., 2007). So, this algorithm is still slow and cannot be used for large graphs (Goyal et al., 2011). Generally, LDAG-based algorithms perform better than optimized CELF algorithms (Chen et al., 2010).

Simpath is also a lazy-forward optimized version of CELF (Goyal et al., 2011b). The algorithm inspects all paths from the initial seeds to maximize influence in the LT model. The problem of computing a simple path is #P-hard (Valiant, 1979). Simpath uses the vertex cover method to reduce the number of iterations of the algorithm and the resulted nodes are inspected for influence maximization. In this method, the number of selected nodes is increased after each iteration of the algorithm, so the influence maximization process becomes very slow. Therefore, further optimization is applied to reduce the run time of the iterations.

More recently, in (Zhou, Zhang, Guo, & Guo, 2014), the authors derive an upper bound for the spread function under the LT model. They propose an efficient UBLF algorithm by incorporating the bound into CELF. Experimental results demonstrate that UBLF, compared with CELF, reduces Monte Carlo simulations and reduces the execution time when the size of seed set is small. In (Narayanam & Narahari, 2010), the authors proposed a Shapley value based heuristic SPIN for the LT model. SPIN evaluates the