Knowledge-Based Systems 000 (2016) 1–12

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Big social network influence maximization via recursively estimating influence spread

Wei-Xue Lu[a], Chuan Zhou[b,d], Jia Wu[c,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[c] Centre for Quantum Computation and Intelligent Systems (QCIS), Faculty of Engineering & Information Technology, University of Technology Sydney, NSW 2007, Australia
[d] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 25 February 2016
Revised 22 September 2016
Accepted 25 September 2016
Available online xxx

Keywords:
Greedy algorithms
Social network
Influence maximization

ABSTRACT

Influence maximization aims to find a set of highly influential nodes in a social network to maximize the spread of influence. Although the problem has been widely studied, it is still challenging to design algorithms to meet three requirements simultaneously, i.e., fast computation, guaranteed accuracy, and low memory consumption that scales well to a big network. Existing heuristic algorithms are scalable but suffer from unguaranteed accuracy. Greedy algorithms such as CELF [1] are accurate with theoretical guarantee but incur heavy simulation cost in calculating the influence spread. Moreover, static greedy algorithms are accurate and sufficiently fast, but they suffer extensive memory cost. In this paper, we present a new algorithm to enable greedy algorithms to perform well in big social network influence maximization. Our algorithm recursively estimates the influence spread using reachable probabilities from node to node. We provide three strategies that integrate memory cost and computing efficiency. Experiments demonstrate the high accuracy of our influence estimation. The proposed algorithm is more than 500 times faster than the CELF algorithm on four real world data sets.

1. Introduction

Social networks [2] are of vital importance in information diffusion [3,4] and viral marketing [5,6], especially when social media is combined with other entities, such as recommendation systems [7,8], or business [9]. Influence maximization in social networks is defined as finding a subset (called *seed set*) of nodes that can trigger the largest number of users propagating the given information. The problem can be formulated as a discrete optimization problem under the independent cascade model (IC) [10,11] and the linear threshold model (LT) [12]. It has been proved NP-hard, and thus many approximation algorithms and heuristic methods have been developed.

Heuristic algorithms, such as DegreeDiscount [13], PMIA [14], IRIE [15], and Group-PageRank [16], are scalable in big networks but not robust with respect to network structures because they cannot guarantee theoretical accuracy.

On the other hand, greedy algorithms [12] are usually favored for obtaining near-optimal solutions as they have theoretical guarantee due to the submodularity of the target influence function. However, conventional speed-up techniques, such as Cost-Effective Lazy Forward strategy (called CELF optimization) [1], MixedGreedy [13], CELF++ [17], UBLF [18], are still not suitable for processing large-scale networks with millions of nodes because of their high time complexity and the unavoidable cost of simulating influence cascades. Thus, some static greedy algorithms, such as StaticGreedy [19], pruned simulation strategy [20] and incremental strategy [21], have been proposed. These methods drastically improve efficiency by using snapshots instead of simulating cascades, but they suffer from severe memory cost, which limits their application to big networks.

In summary, all these methods still suffer from low scalability, low precision, or high memory cost. No related work has consider these three issues simultaneously. These issues mainly lie in the process of simulating the influence spread of any seed set, i.e., if the blackbox, or influence function, is known, the influence maximization problem can be solved within linear time and $O(1)$ space.

In this paper, we focus on resolving the scalability, accuracy, and memory cost dilemma of influence maximization under the independent cascade model and propose a new efficient algorithm for the influence maximization problem through recursively estimating the influence spread.

[*] Corresponding author. Fax: +61 2 9514 2673.
E-mail addresses: luweixue@gmail.com (W.-X. Lu), zhouchuan@iie.ac.cn (C. Zhou), jia.wu@uts.edu.au (J. Wu).

http://dx.doi.org/10.1016/j.knosys.2016.09.020
0950-7051/© 2016 Elsevier B.V. All rights reserved.

Table 1
Major variables used in the paper.

Variables	Descriptions
$G=(V,E)$	Social network G with node set V and edge set E
n, m	The number of nodes and edges respectively
$\sigma(S)$	The expected influence spread of seed set S
$\mathbf{P}(X)$	The distribution of snapshot or instance X of G
$\mathbf{P}_{a\to v}$	Node-to-node probability from node a to node v
$\mathbf{P}_{S\to v}$	The probability of the event that v is reachable from S
$a+b$	Defined as the set of two merged nodes: $\{a\}\cup\{b\}$
N	A positive truncation number
$\mathbf{P}(\cdot\|a\to b)$	The probability of some event under the condition $a\to b$
μ	A measure, i.e. a nonnegative countably additive function
$\mathbf{E}(f)$	The expectation of random variable f
$\vee$, $\wedge$	The supremum and the infimum operator respectively

The main contributions of our paper are:

- We provide three strategies to obtain reachable probabilities from single node to single node (*node-to-node* probabilities) by considering the scalability and memory cost simultaneously.
- We reduce memory cost through keeping only $O(Nn)$ node-to-node probabilities during node selection instead of R subgraphs, which occupy $O(Rm)$ space, generated from the original network. N is a positive constant, m and n are the number of edges and nodes in the network and R is the number of simulations needed to estimate the influence spread.
- We derive a recursive equation to estimate the influence spread of any given seed set and employ the estimation in each round of selecting a new influential node.
- We successfully avoid revisiting generated subgraphs to compute the influence spread and reduce the time complexity of node selection from $O(kRmn)$ to $O(kn^2)$ compared with previous greedy algorithms, where k is the cardinality of the optimal seed set.

Experimental results demonstrate that the proposed influence estimation method can achieve the comparable accuracy with all baseline greedy algorithms. More importantly, in terms of running time, our influence estimation approach performs more than 500 times faster than CELF on four real world datasets.

The rest of the paper is organized as follows. In Section 2, related work for influence maximization is introduced. In Sections 3 and 4, we derive our recursive estimation of the influence spread given any seed set under the assumption that all the probabilities from node to node are already known. In Section 5, three strategies are provided to obtain these probabilities, and in Section 6 we give an overview of our algorithm. In Section 7, experimental results on four real world datasets are presented. Lastly, we conclude the paper in Section 8. Table 1 outlines the major symbols and variables used in the paper.

2. Related work

Many problems, such as viral marketing [5] or outbreak detection [1], can be abstracted as an influence maximization problem. After such abstraction, the corresponding problems can be dealt with from the perspective of designing fast and accurate algorithms. It was Domingos and Richardson [22] who first studied the problem of influence maximization from the algorithmic perspective. Kempe et al. [12] subsequently formulated it as a discrete optimization problem. They proved that the optimization problem is NP-hard, and presented a greedy approximation algorithm (referred to as GeneralGreedy in this paper) which guarantees that the influence spread is within $1-1/e\approx 63\%$ of the optimal result. However, this greedy algorithm is inefficient and not scalable to large scale social networks because a large number of Monte–Carlo simulations are needed to estimate the expected influence spread of each seed set.

Many studies devoted to optimizing Kempe's greedy algorithm by reducing the number of simulations without lowering its solution quality were proposed as a result. The typical algorithm CELF [1] is 700 times faster, but it still takes a few hours for graphs with tens of thousands of nodes. The CELF algorithm was later upgraded to a CELF++ strategy [17], which simultaneously calculates the influence spread for two successive iterations of a greedy algorithm. The NewGreedy algorithm [13] reuses the results of Monte–Carlo simulations to estimate the influence spread for all candidate nodes in each round of choosing a new node to the seed nodes set. Integrating the advantages of both CELF and NewGreedy forms the MixedGreedy algorithm [13]. UBLF [18] drastically reduces the number of simulations in the first round of node selection by introducing an upper bound for every single node. However, these improved greedy algorithms are still inefficient, because they involve too many Monte–Carlo simulations for influence spread estimation.

By reusing the subgraphs (called *snapshots* or *instances* [23]) generated from the original network, static greedy algorithms reduce the number of simulations by two orders of magnitude. StaticGreedy [19] has a speed comparable to some scalable heuristic algorithms, but it is still impractical for large-scale networks to keep these snapshots due to severe memory cost, as [20] has pointed out in their experiments. Ohsaka et al. [20] introduced Pruned BFS to accelerate the estimation of influence spread. The performance of Pruned BFS is better if there is a node whose degree is dominantly larger than that of others, which is not always the case; as stated in [24], the degree distribution is usually heavy-tailed. Lu et al. [21] proposed an incremental strategy to deal with big networks by breaking down the original network into subgraphs and generating simulations on the whole network by joining the results of subgraphs to estimate the influence spread. However, it still suffers from the scalability, accuracy, and memory cost dilemma.

Several heuristics for the independent cascade model have been proposed to avoid using Monte–Carlo simulations. Chen et al. [13] suggested a degree discount heuristic which significantly decreases the running time by only considering the direct influence of a node to its one-hop neighbors. However, this method is restricted to the uniform independent cascade model and the propagation probability must be small enough. PMIA [14] introduces maximum influence paths to estimate the influence spread. Jung et al. [15] proposed the IRIE algorithm which formulates the influence spread using simultaneous linear equations.

Another method of designing heuristics considers the influence maximization problem as a ranking problem. Liu et al. [16] proposed a Group-PageRank strategy based on a similar idea to PageRank. Cheng et al. [25] suggested finding a self-consistent ranking starting with another heuristic method and named it IMRank.

All the above-mentioned methods suffer from low scalability, low precision, or high memory cost issues, and these issues mainly occur in the process of simulating the influence spread of any seed set. If we want to obtain fast and accurate algorithms for the influence maximization problem, we must first estimate the influence spread of any seed set efficiently. To this end, we first carry out the related preliminaries in Section 3, and then introduce our method to estimate the influence spread in Section 4.

3. Preliminaries

3.1. Notations and basic definitions

For similarity, we let $G=(V,E)$ be an undirected network with a node set V of size n and an edge set E of size m. We adopt the