A Comparative Study on Algorithms for Influence Maximization in Social Networks

Extended Abstract

Yu-Hsiang Chung and Tuan-Fang Fan
National Penghu University of Science and Technology
Penghu 880, Taiwan
dffan@npu.edu.tw

Churn-Jung Liau
Academia Sinica
Taipei 115, Taiwan
liaucj@iis.sinica.edu.tw

ABSTRACT

How to disseminate information or ideas through social network connection has received much attention in recent years. The core issue is to find a seed set of initially active individuals that can maximize the influence spread. In this paper, we present a comparative study on three basic algorithms for such issue. Experimental results show that although genetic algorithm can find slightly better solution than other algorithms, it is too time-consuming to be cost-effective. Hence, our on-going work is aimed at improving the search efficiency of different bio-inspired meta-heuristic methods.

CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; *Modeling and simulation*; • **Information systems** → *Social networks*;

KEYWORDS

Social network, influence maximization, greedy algorithm, genetic algorithm, PageRank

ACM Reference Format:
Yu-Hsiang Chung and Tuan-Fang Fan and Churn-Jung Liau. 2018. A Comparative Study on Algorithms for Influence Maximization in Social Networks: Extended Abstract. In *Proceedings of the Genetic and Evolutionary Computation Conference 2018 (GECCO '18 Companion).* ACM, New York, NY, USA, 2 pages. https://doi.org/https://doi.org/10.1145/3205651.3205667

1 INTRODUCTION

With the rapid growth of on-line social networking, the network-based information dissemination has found many applications in human society. A core issue in these applications is how to maximize the diffusion of information which can be formulated as the *influence maximization problem* (IMP)[3–5]. The IMP is to find a subset of initially active individuals who can activate as many as individuals in the social network based on some information propagation models. However, because the IMP is a NP-hard optimization problem [5], there is no theoretical guarantee of optimal solution for existing algorithms. Hence, most algorithms can only claim the near-optimality about their solutions. Consequently, an experimental comparison on the effectiveness and efficiency of different algorithms is needed for the choice of the most appropriate method. In this paper, we present such kind of comparative study on three basic algorithms: greedy method, genetic algorithm, and PageRank, based on the linear-threshold (LT) propagation model.

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/https://doi.org/10.1145/3205651.3205667

2 BASIC DEFINITIONS AND PROBLEM FORMULATION

A social network is comprised of a finite set of individuals and the relationships among the individuals. Formally, a social network $\mathfrak{N} = (V, E, L)$ is a labeled (directed) graph, where V is a set of nodes denoting individuals, E is the set of edges denoting the binary relation, and $L : V \to 2^I$ is a label function such that $L(x) = \{i \in I \mid x \in P_i\}$ with I being the set of possible attributes possessed by individuals.

In general, a social network can induce a *social influence graph* which is defined as a (weighted) directed graph (V, E) endowed with weights on edges in E. We assume that the graph does not have self-loop, i.e., $(x, x) \notin E$ for any $x \in V$. Let $V = \{x_1, \cdots, x_n\}$ and $e = (x_i, x_j) \in E$ be an edge of the graph. Then, the weight $0 \leq p_{ij} \leq 1$ denote the degree of x_i's influence on x_j. For convenience, we can set $p_{ij} = 0$ if $(x_i, x_j) \notin E$. In particular, $p_{ii} = 0$ for any $x_i \in V$. Thus, the set E can be omitted and we can simply represent the influence graph by its weight matrix $\mathbf{p} = [p_{ij}]_{1 \leq i,j \leq n}$, where $n = |V|$. By abusing the terminology somewhat, we also call the matrix $\mathbf{p}$ the influence graph.

From now on, we use $V = \{x_1, \cdots, x_n\}$ and $\mathbf{p}$ to denote the set of individuals in a social network and its associated influence graph respectively. During the process of influence propagation, every individual is either active or inactive. Hence, we denote the state at time t by an $1 \times n$ state vector $\mathbf{s}^t = [s_1^t, \ldots, s_n^t]$, where $s_i^t = 1$ if individual x_i is active at time t, otherwise $s_i^t = 0$. Because we only consider the discrete time model, the time t ranges on the natural numbers $0, 1, 2, \cdots$.

For the LT model, there exists a randomly generated activation threshold $\theta_i^t \in [0, 1]$ beyond which the individual x_i will be influenced at time t. We also use the vector form $\boldsymbol{\theta}^t$ to denote the thresholds of all individuals at time t. In addition, for the LT model, we impose the normality assumption on the weights such that $\sum_{1 \leq i \leq n} p_{ij} = 1$ for any j. Then, in the model, the state transition equation is $\mathbf{s}^{t+1} = \mathbf{s}^t \vee \lceil \mathbf{s}^t \mathbf{p} - \boldsymbol{\theta}^t \rceil$, where $\vee$ and $\lceil \cdot \rceil$ denote the (pointwise) max operation and the ceiling function respectively. In other words, active individuals will remain active forever and an inactive individual x_i will become active at time $t + 1$ if $\sum_{j=1}^{n} (s_j^t \cdot p_{ji}) = \sum_{j : s_j^t = 1} p_{ji} > \theta_i^t$. Intuitively, this means that x_i is activated if the sum of influence weights of his active in-neighbors

is bigger than his threshold. Since the set of individuals is finite, the state transition process will stop in finite iterations. Let t be the smallest time such that $\mathbf{s}^{t+1} = \mathbf{s}^t$. Then, the set of active individuals at time t contains exactly those who are influenced during the diffusion process. However, because $\boldsymbol{\theta}^t$ is a random variable, the final state of the process cannot be determined in a deterministic way. Therefore, we can only compute the expected number of active individuals at the final state. Let $S \subseteq V$ be the set of active individuals, called *seed set*, at time 0. Then, the expected number of active individuals at the final state is called the *influence spread* of S and is denoted by $\sigma(S)$.

Based on the diffusion model, the IMP can be formulated as a kind of optimization problem. That is, the objective of IMP is to find a seed set S that can achieve maximum $\sigma(S)$. However, in many applications, to inject information on members of the seed set is generally not free. Therefore, we have to take the cost of each seed set into consideration. Let $C : V \to \Re^+$ be a cost function on the set of individuals and $B \in \Re^+$ be a budget constraint. Then, the IMP is formulated as the following optimization problem

$$\max\{\sigma(S) \mid S \subseteq V, \sum_{x \in S} C(x) \leq B\}.$$

For the purpose of simplification, it is usually assumed that $C(x) = 1$ for all $x \in V$ and $B = k$ for some positive integer k. Then, the form of the IMP is reduced to $\max_{S \subseteq V, |S| \leq k} \sigma(S)$. In this paper, we will only consider this simplified version of the IMP.

3 ALGORITHMS

As mentioned above, there do not exist efficient algorithms that provide theoretical guarantee to find optimal solution of the IMP problem. Hence, different approximation algorithms and heuristics have been proposed. The most popular ones are the greedy algorithm and its variants. The algorithm starts with an empty seed set and individuals are sequentially added until the cardinality of the seed set is k. At each iteration, we choose, from individuals not in the seed set yet, the individual that can maximally improve the influence spread of the seed set.

Because IMP is an optimization problem, it is expected that some nature-inspired meta-heuristic methods, such as genetic algorithm (GA) can be applied to solving the problem [2].

The main design choices of GA depend on the genetic representation of candidate solutions and the fitness function to evaluate the solution domain. As usual, we can take the objective function of the optimization problem as the fitness function. Hence, in the basic GA for IMP, the fitness function is simply the influence spread function σ. In addition, because each candidate solution is a seed set, its characteristic function can be encoded as a bit vector straightforwardly. However, the usual mutation and crossover operators for bit vectors can not keep the cardinality of a candidate solution fixed. To remedy the problem, we can drop the budget constraint k from the requirement of IMP or modify the applicable genetic operators.

PageRank is an algorithm used by Google search engine to rank web pages by assigning a numerical score to each page to measure its importance [1]. The basic idea is that a page which is linked to by many pages with high scores should receive a higher score itself. This can be applied to IMP by assigning a numerical score to each individual in a social network to reflect its influential power. Then, we can choose the k individuals with highest scores as seeds. Unlike greedy algorithms or GA, the computation of PageRank does not have to run a lot of simulations for influence spread. Instead, the scores arranged in column vector format $bfr = [r_i]_{1 \leq i \leq n}$ can be obtained by solving the following linear equation:

$$\mathbf{r} = \frac{1-d}{n}\mathbf{1} + d\mathbf{pr},$$

where $\mathbf{1} = [1, \cdots, 1]$ is a column vector of length n, and d is called the damping factor and usually set to 0.85. In other words, i's score is larger if he has higher degree of influence (p_{ij}) on high scored (r_j) individuals.

Table 1: Experimental Results

Methods	Running Time (sec.)	Influence Spread
Greedy	147.1387	67.7
GA	12283.182	70.8
PageRank	0.997	70.5

4 EXPERIMENTAL RESULTS AND CONCLUSION

To compare the effectiveness and efficiency of the above-mentioned algorithms, we conduct a series of experiments on a discussion group network with 83 nodes in a local BBS. We assume a budget constraint $k = 3$. In the experiments, we run $R = 10000$ rounds of Monte-Carlo simulations each time when the computation of $\sigma(S)$ is needed. The depth of each simulation round is set to $D = 6$. For the GA implementation, we set the population size as $M = 50$ and the mutation rate as 0.1. We conduct the experiments by running 10 tests for each of the three algorithms and the results are shown in Tables 1. In the case of PageRank, the time for finding the seed set is simply 0.012 seconds. However, we show the total time including the running time for evaluating the influence spread of the seed set. Although, for the small-scaled network used in our experiments, PageRank is the most efficient algorithm and achieves almost the same optimality level as GA, its superiority over other algorithms are not conclusive yet. In the future, we still need to conduct more large-scaled experiments to better understand the performance of different algorithms.

REFERENCES

[1] S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks* 30, 1-7 (1998), 107–117.
[2] D. Bucur and G. Iacca. 2016. *Influence maximization in social networks with genetic algorithms.* Springer, 379–392.
[3] T. Carnes, . Nagarajan, S.M. Wild, and A. van Zuylen. 2007. Maximizing influence in a competitive social network: a follower's perspective. In *Proceedings of the 9th International Conference on Electronic Commerce: The Wireless World of Electronic Commerce*. 351–360.
[4] P.M. Domingos and M. Richardson. 2001. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 57–66.
[5] D. Kempe, J.M. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 137–146.