A Multi-objective Optimization Approach for Influence Maximization in Social Networks

Jian-bin Guo[1], Fu-zan Chen[2(✉)], and Min-qiang Li[1,2]

[1] College of Management and Economics, Tianjin University, Tianjin, China
[2] State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Tianjin, China
fzchen@tju.edu.cn

Abstract. Influence maximization (IM) is to select a set of seed nodes in a social network that maximizes the influence spread. The scalability of IM is a key factor in large scale online social networks. Most of existing approaches, such as greedy approaches and heuristic approaches, are not scalable or don't provide consistently good performance on influence spreads. In this paper, we propose a multi-objective optimization method for IM problem. The IM problem is formulated to a multi-objective problem (MOP) model including two optimization objectives, i.e., spread of influence and cost. Furthermore, we develop a multi-objective differential evolution algorithm to solve the MOP model of the IM problem. Finally, we evaluate the proposed method on a real-world dataset. The experimental results show that the proposed method has a good performance in terms of effectiveness.

Keywords: Influence maximization · Multi-objective differential evolution algorithm · Multi-objective optimization model · Social network

1 Introduction

With the rapid development of social networks, more and more people exchange information on social networks. Social networks provides a broader platform for information propagation. The commercial value embedded in social networks gradually emerges (such as 'virtue marketing').

Influence maximization (IM) problem is a task of finding a set of seed nodes (i.e., seed set) that make these nodes have the broadest influence spread based on a specific propagation model. It has attracted great attention of scholars and industrial practice.

With the increasing expansion of social networks, searching a seed set on a large-scale social network is a NP-hard problem. Most of existing approaches fail to get the optimal solution for a large-scale social network in a reasonable time.

Furthermore, most of previous researches only pursue the maximal of influence spread. Few of them consider cost of spread i.e., payment for delivering targeted information.

To solve the problem of IM, we propose a multi-objective optimization method. The IM problem is formulated to a multi-objective optimization problem (MOP) model

G. Q. Huang et al. (eds.), *Proceeding of the 24th International Conference on Industrial Engineering and Engineering Management 2018*,
https://doi.org/10.1007/978-981-13-3402-3_74

including two optimization objectives, i.e., influence spread and cost. Furthermore, we develop a multi-objective differential evolution algorithm to solve the MOP model of IM. Finally, the proposed approach is validated on a real-world dataset.

The rest of the paper is organized as follows. Section 2 provides a review of related work. In Sect. 3, we present the proposed multi-objective optimization model for the IM problem. Section 4 develops a multi-objective differential evolution algorithm to solve the MOP model of IM. Section 5 reports experimental analysis on a real-world dataset to validate the proposed method. We draw a conclusion and discuss the future work in Sect. 6.

2 Related Work

Existing approaches for the IM can be divided into two types, i.e., greedy approaches and heuristic approaches. However, greedy approaches are not scalable and heuristic approaches do not provide consistently good performance on influence spreads.

Domingos and Richardson firstly considered IM problem as an algorithmic problem, using *Markov random field* modeling to simulate the influence propagation process, and proposed a heuristic method to solve this problem [1].

Kempe et al. considered the influence maximization problem as a discrete optimization problem [2]. His work focused on two propagation models: *IC (independent cascade model)* and *LT (linear threshold model).* Based on the above two propagation models, a greedy algorithm is proposed to solve the IM problem.

In the above work, *Monte Carlo simulation* is used to estimate the propagation range, but the use of *Monte Carlo simulation* increases the running time of the algorithm and makes it difficult to apply to large-scale social networks. Therefore, most of the scholars put a lot of effort into improving the efficiency of the algorithm. Leskovec et al. used the submodularity of propagation to propose that *CELF* algorithm can greatly improve the algorithm running speed [3]. Chen et al. proposed algorithm *CGA* its main idea is to find the optimal set of seed nodes from the newly constructed subsocial network diagram [4]. Goyal et al. proposed a *CELF++* algorithm for further improvement of *CELF* [6].

Some scholars proposed the use of evolutionary algorithms to solve this NP-hard problem. Gong et al. proposes using particle swarm optimization algorithm (PSO) to solve IM problem [11]. Bucur et al. proposes the use of genetic algorithms (GA) to solve IM problem [12]. Jiang Q et al. proposed to use simulated annealing algorithm (SA) to solve IM problem [13]. The use of evolutionary algorithms greatly shortens the time for solving IM problems.

Node influence measure is based on some characteristics of the network to construct the corresponding formula to calculate the node's global influence. Kempe et al. propose to use *Degree Centrality* and *Closeness Centrality* to measure the node's global influence according to the node's influence definition in social networks [1]. Cha et al. proposed using the '*Retweet*', '*Comment*' and '*Mention*' in social networks to measure the influence of nodes in twitter [7]. Romero et al. use the *Hirsch Index* to estimate the global influence of nodes in social networks [8]. Gayo-Avello et al.

proposed a physics-based variable mass system's influence metric [9]. Kitsak et al. proposed a measure of the influence of *k-shell* decomposition on the nodes of influence in dynamic propagation [10].

3 Multi-object Optimization Model for the IM

In practice, enterprises releasing information in social networks usually pay money or give coupons to individuals according to their influence in the social network in order to encourage them to retweet the targeted information. As mentioned, existing literatures of IM pursuit of the maximal spread of influence, but few of them consider spread cost, i.e., payment or coupons. Besides of the spread of influence, we consider the spread cost are considered. The problem of IM is formulated as a MOP and a multi-objective optimization model is proposed in this section.

3.1 Spread of Influence

The *independent cascade (IC) model* is used in this paper to simulate the propagation process of the targeted information [2]. Suppose a given social network is represented by a directed graph $G = (V, E, P)$. The nodes $v \in V$ in the directed graph represent the users in the social network, the edges $(u, v) \in E$ represent the relationships between users, and the weights P_{uv} on the edges (u, v) represent the probability of influence between users.

There are a number of calculation methods defined for the evaluation of the probability of activation between nodes. Although the traditional heuristic calculation method is simple to calculate, the final probability is difficult to fit the true propagation probability.

Zhang et al. proposes a method for calculating the probability of influence between nodes [14], which considers that the more frequent the interaction between two nodes, the more likely the two nodes will influence each other, and the greater the probability of activation between nodes. They define that the activation probability of node u to activate node v is P_{uv}. ω_{uv} denotes the weight on the edge from u to v. $p \in [0, 1]$ is a designated propagation probability. The probability of activation between nodes is defined as:

$$p_{uv} = 1 - (1 - p)^{\omega_{uv}} \tag{1}$$

The probability of influence between nodes calculated by the above method are more fitting the actual propagation probabilities than the heuristic activation probabilities of the nodes in the universal *IC* model, so that the actual propagation of information can be more accurately simulated.

The *independent cascade model (IC)* simulates the random propagation of information. We represent the social network as a directed graph $G = (V, E, P)$. The nodes v contained in V represent users in the social network. The edges e contained in E represent the relationships between nodes and are defined for each edge in E. P_{uv} represents the influence probability of the edge (u, v).

The state of the node in this model includes two types: active, inactive, and some nodes (called seed nodes) are pre-activated at the initial stage $t = 0$ to form a node set S. In any step $t > 1$, if node u has been activated at step $t - 1$, then node u has only one chance to attempt to activate its inactivated neighbor node v with probability P_{uv}, and the node cannot be activated once it is activated. The process terminates when no new node is activated.

3.2 Cost of Influence

In practice, discounts or coupons issued by companies in order to encourage initial users to deliver targeted information are the major components of the cost of information spread.

And the greater the influence of the initial user, the greater the cost of activating the user. Therefore, in this paper we assume that the user's activation cost is positively related to the user's influence.

So we define the activation cost of the seed node based on the global influence of the node itself. The cost function is defined as follows:

$$\cos t(S) = \sum_{u \in S} \sum_{v \in V} p_{uv} \tag{2}$$

S represents the seed set, and P_{uv} represents the probability that node u activates node v.

We define the sum of the activation probabilities between u and all its *outgoing degree node* as the global influence of node u on the social network.

3.3 Multi-objective Optimization Model for the IM

In this paper, we consider the two objectives of maximizing the influence spread and minimizing the cost of influence, and propose a multi-objective optimization based influence maximization model. Two objective function formulas are defined as follows:

(1) Maximization of Influence spread

We use the *IC* model to simulate the final number of nodes activated by the seed set as the final influence spread. Previous scheme often used *Monte Carlo simulations* to simulate the final range of propagation, but multiple *Monte Carlo* simulations were very time- consuming, so we used a more simple *LIE* function to calculate the number of nodes what were finally activated [11].

$$S^* = \arg \max_{|S|=k, S \in V} LIE(S) \tag{3}$$

S represents the selected seed set, and k represents the number of seed nodes.

(2) Minimization of cost

Minimizing the cost of influence is equivalent to minimizing the activation cost of the initial set of nodes. Therefore, the objective function is to minimize the activation cost of the initial seed set. So cost minimization objective function:

$$S^* = \arg \max_{|S|=k, S \in V} \cos t(S) \tag{4}$$

3.4 Mathematical Model of Multi-objective Optimization

Because the MOP model is suitable for solving the minimization problem. And our objective function contains a maximum spread of influence. So we need to convert the

objective function. D represents the number of nodes in the social network. So we turn the optimization goal of the maximum spread of influence into the minimum number of inactive nodes.

$$\begin{cases} f_1(S) = D - LIE(S) \\ f_2(S) = \cos t(S) \end{cases} \tag{5}$$

The multi-objective optimization mathematical model constructed in this paper is as follows:

$$\begin{aligned} &\min(f(S) = \min(f_1(S), f_2(S)) \\ &s.t.\ |S| = k; \\ &S \in V \end{aligned} \tag{6}$$

4 Solving Algorithm for the MOP Model

In this section, we develop a multi-objective differential evolutionary algorithm (*MODEA*) to solve the MOP model for IM.

For the above multi-objective optimization model, we adopt the multi-objective evolutionary algorithm to optimize the solution. In view of the complexity of the problem of maximizing influence, we have improved the optimization algorithm so that it can converge more quickly to the Pareto frontier.

Differential evolution algorithm is an evolutionary algorithm based on evolutionary theory of genetic algorithm. The essence is a multi-objective optimization algorithm (*MOEAs*) for solving the global optimal solution in multi-dimensional space. Has the advantages of easy to use, simple structure, fast convergence, and robustness. We extend the DE algorithm to a multi-objective form (*MODEA*) and improve the selection strategy of the DE algorithm so that it can quickly converge to the Pareto frontier under the premise of ensuring the diversity of the optimal solution distribution.

The principles in *DE* were simplicity, efficiency, and the use of floating-point encoding instead of binary numbers. As with traditional evolution, the *DE* algorithm owns an initial population and is promoted by selection, mutation and crossover during the iteration.

4.1 Population Initialization

In the solution space, N individuals are randomly generated, each of which is represented as n-dimensional vectors such as:

$$X_i(0) = (X_{i1}(0), X_{i2}(0), \ldots, X_{iN}(0)), i = 1, 2, 3, \ldots, N; \tag{7}$$

The i th individual's j th dimension value is as follows:

$$\begin{aligned} &X_{ij}(0) = L_{j-min} + rand(0,1) * \left(L_{j-max} - L_{j-min}\right) \\ &i = 1, 2, 3, \ldots, N; \\ &j = 1, 2, 3, \ldots, k; \end{aligned} \tag{8}$$

$L_{j\text{-}min}$ represents the minimum value on the j-th dimension, and $L_{j\text{-}max}$ represents the maximum value in the j-th dimension.

4.2 Mutation

In the gth iteration, 3 individuals were randomly selected from the population $X_{p1}(g)$, $X_{p2}(g)$, $X_{p3}(g)$ and p1 ≠ p2 ≠ p3. The resulting mutation vector is:

$$H_i(g) = X_{p1}(g) + F * \left(X_{p2}(g) - X_{p3}(g)\right) \tag{9}$$

where $X_{p2}(g) - X_{p3}(g)$ is the difference vector, F is the differential weight, for the differential weight F, generally choose between [0, 2], usually take 0.5.

4.3 Crossover

$$V_{i,j} = \begin{cases} h_{i,j}(g), rand(0,1) \leq cr \\ X_{i,j}(g), else \end{cases} \tag{10}$$

where $cr \in [0, 1]$ is the crossover probability.

4.4 Selection

To expand *DE* algorithm into multi-objective optimization algorithm, we need to improve the selection operation in *DE* algorithm.

Selection in *MODE*A is based on the following rules:

(1) When there is a Pareto dominance relationship between two solution vectors, we choose a better solution vector based on Pareto dominance to enter the next generation population.
(2) When there is no Pareto dominance relationship between two solution vectors, we choose all the two solution vectors into the next generation population.

After an iteration, the size of the population may increase. If the population size grows to a pre-set threshold, we use a selection operation similar to the one in *NSGA-II* to resize the population to the original size. The solution vector is sorted according to the indexes of non-domination and crowding degree, and delete the solution vectors with poor performance to reduce the population size to the initial size.

The increment ε for each iteration size of the population is between N and $2N$ (N is assumed to be the initial population size), since both vectors can enter the next generation without any dominance between the vectors. We assume that the population size threshold is $2N$. When the population size increases to $2N$ after many times of iterations, a selection operation with non-dominated sorting is called to reduce the population size to the original population size.

From the above process, we know that the improved algorithm does not have to call selection operations with non-dominated sorting every time, and the number of invocations is much smaller than that of the *NSGA-II* algorithm. However, the time

complexity of a select operation with non-dominated sorting is $O(g * N^2)$, which is a major part of the computational complexity of the algorithm. Reducing the number of calls can significantly shorten the algorithm's running time. In the improved algorithm, the increase of the population size within the preset threshold makes the algorithm not easy to be trapped in the local optimum, which is beneficial to increase the diversity of the optimal solution. The pseudocode of the above algorithm is as follows:

Algorithm 1. MODEA

Input: social graph $G=(V,E,P)$, M *and* $n \in N^{+}$, cr=0.2, population size threshold: T .
Output: a series of seed set $S=\{s_1,s_2,\ldots,s_M\}$
For i =1,2,…,M:
  For j=1,2,…,n:
    $X_{i,j}(0) \leftarrow L_{j\_min}+rand(0,1)*(L_{j\text{-}max} - L_{j\text{-}min})$
  end
  $X_i(0) \leftarrow \{X_{i,1}(0), X_{i,2}(0),\ldots,X_{i,n}(0)\}$
end
$X \leftarrow \{X_1(0), X_2(0),\ldots, X_M(0)\}$
Repeat:
  For i =1,2,…,M:
    $\{p_1,p_2,p_3\} \leftarrow rand(1,M)$ // Randomly select three different integers from 1 to M.
    $H_i(g) \leftarrow X_{p1}(g)+F*[X_{p2}(g) - X_{p3}(g)]$
  end
  For i=1,2,…,M:
    For j=1,2,…,n:
      If rand(0,1)<cr:
        $V_{i,j}(g) \leftarrow H_{i,j}(g)$
      else:
        $V_{i,j}(g) \leftarrow X_{i,j}(g)$
    end
  end
  For i=1,2,…,M:
    if f $[V_i(g)]$ <f $[X_i(g)]$:
      $X_i(g+1) \leftarrow V_i(g)$
    else:
      $X_i(g+1) \leftarrow V_i(g)$
      $X_{i+1}(g+1) \leftarrow X_{i+1}(g)$
  end
  If pop[X(g+1)]>T:
    Non-dominant sorting[X(g+1)]
  else:
    pass
Until convergence

5 Experimental Analysis

In this section, we construct several experiments on a public real-world dataset used in [5] to evaluate the performance of the proposed method.

Dataset is a collection of data from academics and scholarly works obtained from the *ArnetMiner* Academic Search System. The collection contains 2,162 academic authors and 2,555 scholarly articles. And based on academic articles to establish a reference of 19,875 times between scholars.

We argue that the more times Scholar *A* refers to Scholar *B* in a certain field, the more influence Scholar *B* has on Scholar *A* in a certain field, which means Scholar *B* can activate Scholar *A* with a greater probability in this field.

We use real-coded instead of binary code. Each population individual represents a set of candidate seed nodes. The coding of the individual population consists of the coding of the nodes it contains.

In experiments, the weights of the edges between nodes in the network represent the number of interaction between the nodes. The default parameter p has a value of 0.5.

For our proposed new multi-objective differential algorithm, we compared the performance of the algorithm with the most commonly used *NSGA-II* algorithm for solving multi-objective optimization problems. In the experiment, the population size $N = 30$, the number of iterations $g = 300$, the size of the seed node set $k = 20$, and the crossover probability $cr = 0.2$. The final experiment result are Fig. 1. From the experiment result, we can see that the new proposed multi-objective differential algorithm can obtain more excellent frontiers than the NSGA-II algorithm.

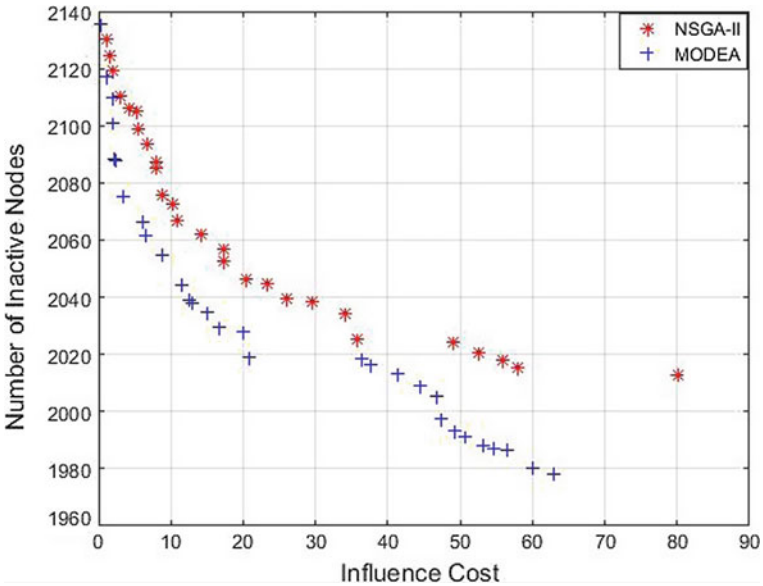

Fig. 1. Comparison of NSGA-II and MODEA

From the above figure, we can obtain a smooth frontier surface by using the multi-objective optimization model to solve the problem of maximizing the influence. Increased diversity of feasible solutions, closer to actual business analysis, providing more direct and comprehensive information for business decisions.

6 Conclusion

The influence maximization problem based on multi-objective optimization proposed in this paper can quickly and accurately (using improved multi-objective evolutionary algorithm) find the optimal set of initial seed nodes for particular targeted information so that the final targeted information has the largest spread of influence. This not only saves costs but also enables more intuitively support for actual decisions and reduces the difficulty of applying research to practical business.

However, there are some limitations to the solution proposed in this paper. For example, the cost function constructed in this paper is only a linear function of the global influence of the node. In reality, the effect of fitting the actual cost may not be very good. The introduction of a more accurate pricing model is an exploration of our future work.

Acknowledgements. The work was supported by the Key Program of National Natural Science Foundation of China (No. 71631003) and the General Program of National Natural Science Foundation of China (No. 71771169).

References

1. P. Domingos, M. Richardson, Mining the network value of customers, in *KDD* (2001), pp. 57–66
2. D. Kempe, J. Kleinberg, Maximizing the spread of influence through a social network, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03 (ACM, 2003), pp. 137–146
3. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J.M. Van Briesen, N.S. Glance, Cost-effective outbreak detection in networks, in *KDD* (2007), pp. 420–429
4. W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in *KDD* (2009), pp. 199–208
5. J. Tang, J. Sun, C. Wang et al., Social influence analysis in large-scale networks, in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2009), pp. 807–816
6. A. Goyal, W. Lu, L.V. Lakshmanan, Celf++: optimizing the greedy algorithm for influence maximization in social networks, in *Proceedings of the 20th International Conference Companion on World Wide Web* (ACM, 2011), pp. 47–48
7. M. Cha, H. Haddadi, F. Benevenuto, P.K. Gummadi, Measuring user influence in twitter: the million follower fallacy, in *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ed. by W.W. Cohen, S. Gosling, ICWSM 2010, Washington, DC, USA, May 23–26, 2010 (The AAAI Press, 2010)

8. D.M. Romero, W. Galuba, S. Asur et al., Influence and passivity in social media, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, Berlin, Heidelberg, 2011), pp. 113–114
9. D. Gayo-Avello, D.J. Brenes, D. Fernández-Fernández, M.E. Fernández-Menéndez, R. García-Suárez, De retibus socialibus et legibus momenti. EPL (Europhysics Letters) **94**(3), 38001 (2011)
10. M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks. Nat. Phys. **6**, 888–893 (2010)
11. M. Gong, J. Yan, B. Shen et al., Influence maximization in social networks based on discrete particle swarm optimization. Inf. Sci. **367**(C), 600–614 (2016)
12. D. Bucur, G. Iacca, Influence maximization in social networks with genetic algorithms, in *European Conference on the Applications of Evolutionary Computation* (Springer, Cham, 2016), pp. 379–392
13. Q. Jiang, G. Song, G. Cong et al., Simulated annealing based influence maximization in social networks, in *AAAI Conference on Artificial Intelligence* (AAAI Press, 2011), pp. 127–132
14. X. Zhang, J. Zhu, Q. Wang et al., Identifying influential nodes in complex networks with community structure. Knowl. Based Syst. **42**(2), 74–84 (2013)