Active Reverse Path Based Reverse Influence Maximization in Social Networks

Ashis Talukder, and Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University, South Korea
Email: {ashis, cshong}@khu.ac.kr

Abstract

Reverse Influence Maximization (RIM) is a new research direction in the influence maximization problem domain which deals with seeding cost of viral marketing in social networks. The influence maximization (IM) searches for a small seed set that maximizes the spread of influence in the network, measured by the number of nodes activated by the seed set. On the other hand, the RIM finds the optimized cost of activating the seed nodes and the cost is measured by the number of nodes that need to be activated in order to activate the seed nodes. In this paper, we propose an Active Reverse Path-based model (ARP-RIM) which jointly employs the Independent Cascade (IC) model, voting model, and greedy approach to solve the RIM problem. The ARP-RIM model meets the challenges of the RIM problem more efficiently. We simulate our model with two real datasets of two popular social networks and the result shows that the ARP-RIM model outperforms the existing models.

1. Introduction

Social networks play a vital role in originating and disseminating information. As a result, social networks have gained great emphasize in the domain of business and research as well. The Influence Maximization (IM) is an approach to estimate a small seed set that maximies the spread of influence in the network [1].

The Reverse Influence Maximization (RIM), on the other hand, working in the reverse direction as compared to the traditional IM problem, calculates the seeding cost of viral marketing [2] which is returned by the minimum number of nodes that must be activated in order to activate the seed users. Talukder et al. [2] proposed two random models, one of which is an extension of the Linear threshold model [1]. Likewise the IM, the RIM can be applied in attractive applications $e.g.$ profit maximization, rumor detection [3], expert search etc.

The major challenges of the RIM problem include setting stopping criteria, handling three basic network structures (BNC), insufficient influence etc. To meet the challenges in a better way, we propose an Active Reverse Path-based (ARP-RIM) model which applies joint Independent Cascade (IC) model [1], and Voting model [4], and greedy approach.

2. Literature Review

The seminal work on the IM research has been carried out by Kempe et al. [1] in 2003. They have formulated the most popular and widely studied Linear Threshold (LT) model and Independent Cascade (IC) model. A heuristic algorithm, Cost-Effective Lazy Forward (CELF) is proposed for outbreak detection in social networks [5]. The CELF outperforms many existing greedy models. Goyal et al. [6] propose an efficient improvement to the CELF model. Chen et al. [7] propose a degree discount heuristic approach that has better results than many of the IM algorithms in terms of running time and efficiency.

The RIM problem was first formulated by Talukder et al. in their research [2]. They propose two solution models of the RIM problem: one is a purely random model (R-RIM) and another is a random extension of LT (RLT-RIM) model.

However, neither of the models meet the challenges of the RIM problem properly. Therefore, we propose an Active Reverse Path-based solution (ARP-RIM) to the RIM problem that handles the challenging issues more effectively.

3. Problem Formulation

Consider a social network represented by a directed graph $G(V, E)$ with $n = |V|$ users and $m = |E|$ social ties among them. For each node $v \in V$, we define two sets D_v^{in} and D_v^{out} as in-neighbors and out-neighbors set with associated indegree and outdegree d_v^{in} and d_v^{out} respectively.

The seed set, S ($k = |S|$), of the IM problem is considered to be the target set in the RIM problem [2]. The RIM problem estimates the seeding cost, $\gamma(S)$, which is given by the minimum number of nodes that must be activated in order to activate the k target nodes.

Definition 1. RIM Problem: Given a social network $G(V, E)$ and a target set S of size k, the RIM problem estimates the seeding cost, $\gamma(S)$, which is defined by the minimum number of nodes that must be activated in order to activate all the nodes in S. □

4. Active Reverse Path (ARP-RIM) Model

In this section, we formulate the proposed Active Reverse Path based model to solve the RIM problem.

A. Active Reverse Path

In order to estimate the marginal cost, $\gamma(v)$, of a node v, we consider all the reverse paths $R_u = \{r_{u1}, r_{u2}, \dots, r_{ul}\}$, starting at the each $u \in D_v^{in}$.

Definition 2. Reverse Path (RP): If a node sequence, $P_{u_1} = \{u_1 \rightsquigarrow u_2 \rightsquigarrow \cdots \rightsquigarrow u_p\}$ is a path starting at u_1 in G, then $R_{u_p} = \{u_p \rightsquigarrow u_{p-1} \rightsquigarrow \cdots \rightsquigarrow u_1\}$ is a reverse path starting at u_p in G. □

Algorithm 1: ARP-RIM Model

```
Input: $G(V, E), S$
Result: $\Gamma(S), \gamma(S)$
$Q = \emptyset$, $\Gamma(u) = \emptyset$; /* Eestimating $\Gamma(u)$, $u \in D_v^{in}$ */
for each $u \in D_v^{in}$ do
    $InsertQ(u)$;
    $\Gamma(u) = \Gamma(u) \cup \{u\}$;
    while $Q \neq \emptyset$ do
        if $u$ is activated by at least one $w \in D_u^{in}$ with
          probablity $p \in \{0.001, 0.01, 0.1\}$ then
            $InsertQ(u)$;
        end
        $\Gamma(u) = \Gamma(u) \cup \{u\}$;
    end
end
$\Gamma(v) = \emptyset$;        /* Eestimating $\Gamma(v)$, $v \in S$ */
for $i = 1$ to $\lfloor \frac{1}{2} d_v^{in} \rfloor + 1$ do
    $u = \underset{u \in D_v^{in}}{arg\ min} \left| [\Gamma(v) \cup \{u\}] - \Gamma v \right|$;
    $\Gamma(v) = \Gamma(v) \cup \{u\}$;
    $D_v^{in} = D_v^{in} - \{u\}$;
end
$\Gamma(S) = \emptyset$;        /* Eestimating $\gamma(S)$ */
for $v \in S$ do
    $\Gamma(S) = \Gamma(S) \cup \Gamma(v)$;
end
$\gamma(S) = |\Gamma(S)|$;
return $\gamma(v)$;
```

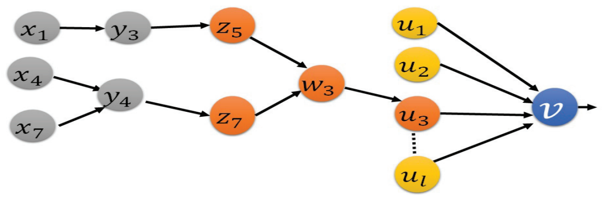


Figure 1: The RPs starting at u_3 are $R_{u_3} = \{\{u_3 \rightsquigarrow w_3 \rightsquigarrow z_5 \rightsquigarrow y_3 \rightsquigarrow x_1\}, \{u_3 \rightsquigarrow w_3 \rightsquigarrow z_7 \rightsquigarrow y_4 \rightsquigarrow x_4\}, \{u_3 \rightsquigarrow w_3 \rightsquigarrow z_7 \rightsquigarrow y_4 \rightsquigarrow x_7\}\}$. Let us assume that the nodes u_3, w_3, z_5 and z_7 are activated by the IC model but y_3 and y_4 are not. Since y_3 and y_4 are not activated, rest part of the RPs (x_1, x_4, x_7) are ignored. Hence, the ARPs starting at u_3 are $AR_{u_3} = \{\{u_3 \rightsquigarrow w_3 \rightsquigarrow z_5\}, \{u_3 \rightsquigarrow w_3 \rightsquigarrow z_7\}\}$. Therefore, $\Gamma(u_3) = \{u_3, w_3, z_5, z_7\}$ and $\gamma(u_3) = 4$.

Definition 3. Active Reverse Path (ARP): If a node sequence, $R_{u_1} = \{u_1 \rightsquigarrow u_2 \rightsquigarrow \cdots \rightsquigarrow u_{i-1} \rightsquigarrow u_i \rightsquigarrow u_{i+1} \cdots \rightsquigarrow u_l\}$ is a reverse path in G and the node u_i is not activated in the activation process (by IC technique), then $AR_{u_1} = \{u_1 \rightsquigarrow u_2 \rightsquigarrow \cdots \rightsquigarrow u_{i-1}\}$ is the associated Active Reverse Path. □

We find all the ARP starting at u_i, $AR_{u_i} = \{ar_{u1}, ar_{u2}, \ldots, ar_{ul}\}$, for all the reverse paths, using the IC technique applied in reverse order. Every inactive node on the reverse path is given a single chance to be activated by its in-neighbors, using a biased coin toss with probability p as stated in the Fig. 1. This probability may be a constant value or can be calculated by weighted cascade model [1], [8], [9]. However, we select the value of p using the Tri-valency model [8], [9], [10].

Now, we compute the cost of each ARP, ar_{ui}, which is measured by the number of nodes in the path as shown in the Eq. 1.

$$\Gamma(ar_{ui}) = \{w | w \in ar_{ui}\} \tag{1}$$

In the following step, we find the seeding cost of all the in-neighbors u of node v by combining the seeding costs of all the ARPs starting from u by the following equation:

$$\Gamma(R_u) = \cup_{ar_{ui} \in R_u} \Gamma(ar_{ui}) \tag{2}$$

B. Marginal Seeding Cost

Now, we have the target node v and all its in-neighbors $D_v^{in} = \{u_1, u_2, ..., u_l\}$ and their associated seeding cost set $\Gamma(u_i)$ as depicted in the Fig. 2.

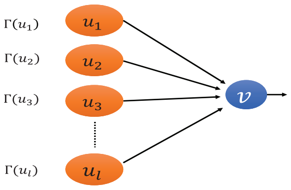


Figure 2: Target node activation by voting.

According to the voting method, we select $\lfloor \frac{1}{2} d_v^{in} \rfloor + 1$ nodes from the set D_v^{in} to get $\gamma(v)$, such that the aggregated seeding cost of the selected nodes is minimized as expressed in the Eq. 3.

$$\Gamma(v) = \underset{|\Gamma(v)| = \lfloor \frac{1}{2} d_v^{in} \rfloor + 1}{arg\ min} \left| \left[\cup_{u_i \in D_v^{in}} \Gamma(u_i) \right] \cup \{v\} \right| \tag{3}$$

We apply the greedy approach to choose the nodes instead of subset problem and this improves the running time by reducing the exponential time problem to a linear time problem.

C. Seeding Cost of Target Set

The seeding cost set $\Gamma(S)$ is computed by combining the marginal seeding cost sets of all $v \in S$ and is given by:

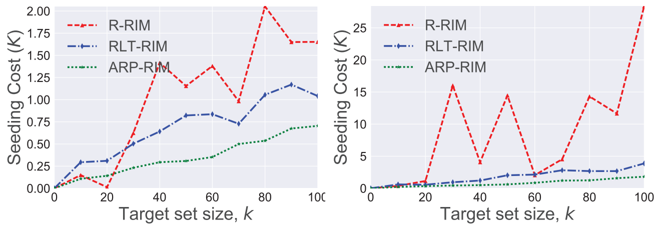
$$\Gamma(S) = \left[\cup_{v \in S} \Gamma(v) \right] \tag{4}$$

The final seeding cost $\gamma(S)$ of the target set S is given by:

$$\gamma(S) = |\Gamma(S)| \tag{5}$$

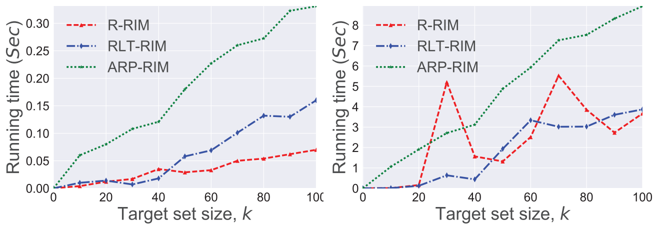
Theorem 1. *The RIM problem is NP-Hard.*

Proof. The Knapsack problem can be reduced to the RIM problem defined in the equations (1) to (5). The Knapsack problem is NP-Hard [2] and therefore, RIM problem is NP-Hard as well. □



(a) Facebook data. (b) Twitter data.

Figure 3: The seeding cost for different k values.



(a) Facebook data. (b) Twitter data.

Figure 4: The running time (in sec) for different k values.

D. *The ARP-RIM algorithm*

The ARP-RIM algorithm, as stated in the Alg. 1, finds the seeding cost of all in-neighbors of all target nodes (lines 1 to 10) firstly. Then, it calculates the optimized marginal seeding cost set, $\Gamma(v)$, $\forall v \in S$ by voting technique (lines 12 to 17) and finally, it combines all marginal costs to estimate the desired seeding cost, $\gamma(S)$ by lines from 18 to 23.

All the ARPs can be computed in $O(n+m)$ time by a breadth or depth first search and all the marginal seeding cost, $\Gamma(v)$ can be computed in $O(d)$ time by the greedy method, where d = maximum degree in G. Therefore, the proposed algorithm has complexity $O(kd(n+m))$.

5. Performance Evaluation

We evaluate the performance of the proposed ARP-RIM model using Pyhton programs and applying Monte Carlo (MC) simulation [1]. By executing the programs $10,000$ times on both the datasets: Facebook[1] and Twitter[2] (see Table I), we take the average values of all the parameters.

Table I: Dataset description

Networks	ego-Facebook	ego-Twitter
Nodes	$4,039$	$81,306$
Edges	$88,234$	$1,768,149$

The Fig. 3 depicts that the proposed ARP-RIM has lower seeding cost than any of the R-RIM and RLT-RIM algorithms. Moreover, the existing models have drastic fluctuation in the result due to the random nature but our algorithm does not suffer from this flaws. Again, by applying the IC model, the ARP-RIM model resolves the issue of stopping criteria which is a major drawback of the existing algorithms [2].

On the other hand, the running time of our algorithm is $O(kd(m+n))$ which is higher than that of the existing models which having running time $O(kd^2)$. The experimental results also reveal the same fact as shown in the Fig. 4. This is due to the higher running time to calculate the ARPs. It contributes the major part, $O(m+n)$, to the running time of the proposed ARP-RIM algorithm. However, in spite of having little higher running time, our algorithm gives better seeding cost and meets the challenges more efficiently.

6. Conclusion

In this paper, we introduce an Active Reverse Path-based solution (ARP-RIM) to the RIM problem to find the seeding cost of viral marketing in the social networks. The ARP-RIM model jointly employs Independent Cascade (IC) model and Voting model along with greedy optimization. The proposed model resolves challenging issues of the RIM problem more efficiently and the experimental results show that it outperforms the existing algorithms though it consumes little more time.

***Acknowledgements*—**This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2017R1A2A2A05000995).
*Dr. CS Hong is the corresponding author.

References

[1] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[2] A. Talukder, M. G. R. Alam, A. K. Bairagi, S. F. Abedin, M. A. Layek, H. T. Nguyen, and C. S. Hong, “A cost optimized influence maximization in social networks,” in *2017 IEEE The 19th Asia-Pacific Network Operations and Management Symposium (APNOMS 2017)*. IEEE, 2017.

[3] A. Talukder, , R. Kalam, A. K. Bairagi, M. G. R. Alam, , S. F. Abedin, M. A. Layek, H. T. Nguye, and C. S. Hong, “Rumors in the social network: Finding the offenders using influence maximization,” in *Korean Computer Congress (KCC)*. KCC, 2015, pp. 1214–1216.

[4] H. Zhang, S. Mishra, M. T. Thai, J. Wu, and Y. Wang, “Recent advances in information diffusion and influence maximization in complex social networks,” *Opportunistic Mobile Social Networks*, vol. 37, no. 1.1, 2014.

[5] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 420–429.

[6] A. Goyal, W. Lu, and 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] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 199–208.

[8] A. Arora, S. Galhotra, and S. Ranu, “Debunking the myths of influence maximization: An in-depth benchmarking study,” in *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 2017, pp. 651–666.

[9] W. Chen, C. Wang, and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1029–1038.

[10] K. Jung, W. Heo, and W. Chen, “Irie: Scalable and robust influence maximization in social networks,” in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 918–923.

[1] https://snap.stanford.edu/data/egonets-Facebook.html
[2] https://snap.stanford.edu/data/egonets-Twitter.html