I J R S S I S, Vol. (IX), Issue (1) Jan 2023: 509-514
A Double-Blind Peer Reviewed Journal
OPEN ACCESS
e-ISSN 2347 – 517X
Original Article

INTERNATIONAL JOURNAL OF RESEARCHES IN SOCIAL SCIENCE AND INFORMATION STUDIES
© VMS RESEARCH FOUNDATION www.ijrssis.in

INFLUENCE MAXIMIZATION IN SOCIAL NETWORKS

Pragya V. Tripathi
Department of English, MahilaMahavidyalayaGadchiroli.
*Corresponding Author: tripathipragya2380@gmail.com

Communicated : 20.12.2022 | Revision : 08.01.2023 Accepted : 20.01.2023 | Published: 30.01.2023

ABSTRACT:

The study of viral marketing strategies on social networks has become an area of significant research interest. Influence maximization is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence. The seed nodes are used for viral marketing to gain the maximum profits through the effective word-of-mouth. However, in more real-world cases, marketers usually target certain products at particular groups of customers. Hence using the labeled influence maximization problem, we aim to find a set of seed nodes that can trigger the maximum spread of influence on the target customers in a labeled social network.

The experiment was conducted to evaluate the effectiveness and efficiency of the methods for the proposed labeled influence maximization problem. We employ the Internet Movie Database (IMDb) to construct the labeled social network. Actors and actresses of movies during 1994-1995 are collected and are considered as nodes in the network. The experimental results on IMDb labeled social network show our methods can achieve promising performances on both effectiveness and efficiency.

Key Words:- algorithms, Internet Movie Database, social network, Facebook

INTRODUCTION :

Over the time, many large-scale online social network sites, such as Facebook and Friendster, become successful because they are very effective tools in connecting people and bringing small and disconnected offline social networks together. Moreover, they are also becoming a huge dissemination and marketing platform, allowing information and ideas to influence a large population in a short period of time. However, to fully utilize these social networks as marketing and information dissemination platforms, many challenges have to be met. In this paper, we present our work towards addressing one of the challenges, namely finding influential individuals for maximum spread of influence in a large-scale social network.

This problem, referred to as influence maximization, would be of interest to many companies as well as individuals that want to promote their products, services, and innovative ideas through the powerful word-of-mouth effect. Online social networks provide good opportunities to address this problem, because they are connecting a huge number of people and they collect a huge amount of information about the social network structures and communication dynamics. However, they also present challenges to solve the problem. The social networks are large-scale, have complex connection structures, and are also very dynamic, which means that the solution to the problem needs to be very efficient and scalable.

In this paper, both the target customers and diverse profits of advertised products are considered into the selection of seed individuals in a social network and propose the labeled influence maximization problem in which the label information of each individual is used to model the ideas of target customers and profits of advertised products. The label information is widely available over current social networking services, where people use some attributed labels to describe personal interests, graduated colleges, hometown, age, skills, favorites, etc. Therefore, given (a) a social network, in which each node possesses a set of labels and each edge is associated with a weight of influence

probability, (b) a set of target labels, (c) a designated profit value for each target label, and (d) the budget in terms of number of selected seed nodes, the labeled influence maximization problem is to find a small set of individuals whose spread of influence in a network can make the total profit as more as possible.

Figure 1 illustrates the idea of the proposed labeled influence maximization. There are 11 nodes and 2 labels {A, B} in the social network. Let the budget of seed nodes is set to be 1 and the influence probability on each edge is the same. For the original influence maximization problem that considers only the number of successfully influenced nodes and neglects the labels on nodes, several effective algorithms (e.g. the degree heuristics and the greedy method) will select the node v1 as the seed to have maximum influence spread. On the other hand, for the proposed labeled influence maximization problem, given the target label B, the best seed node is v7 which has the maximum probability to affect nodes labeled B. An intuitive method to find the effective seed node for labeled influence maximization is to remove the label-A nodes and use existing methods for unlabeled ones to select seeds.

However, such approach suffers from two drawbacks. First, the network will become disconnected, and thus the cascades of influence are constrained to happen locally and cannot propagate from one component to another. Second, such removing could delete the best seeds from the network because the best ones could be those without targeted labels.

To solve the labeled influence maximization problem under the independent cascade model, we propose three approaches, which can be divided into two categories. The first category is developed based on existing algorithms for the original influence maximization by considering the label information. Two methods, Labeled New Greedy and Labeled Degree Discount, are modified to consider the targeted labels and profits. The second category is to provide the interactive mechanisms which provide the capability for marketers to plan and evaluate the advertising strategies in a real-time manner. Ideally, the labeled social network can be obtained before marketers plan the strategies, and we can offline find the seed individuals for all kinds of combinations of target labels in an exhaustive way so that the marketers can wait no time to online evaluate their advertising strategies for different products. However, such enumeration of all possible seed nodes is infeasible because of tremendous combinations of target labels and profits. Therefore, we propose the method, Maximum Coverage, which compute influence potentials between individuals offline and finding the top-k seed nodes online for query (target) labels and respective profits. The influence potentials are measured based on the proximities between nodes in the social network.

Such that under the influence cascade model, the expected number of vertices influenced by the k seeds (referred to as the influence spread in the paper) is the largest possible.

Several recent studies aimed at addressing this efficiency issue. Kimura and Saito propose shortest-path based influence cascade models and provide efficient algorithms of compute influence spread under these models. However, since the influence cascade models are different, they do not directly address the efficiency issue of the greedy algorithms for the cascade models. Leskovec et al. propose to select a set of social sensors such that their placements can efficiently detect the propagation of information or virus in a social network. Lappas et al. propose to find a set of effectors who can cause an activation pattern as similar as possible to the given active nodes in a social network. Chen et al. consider negative opinions (due to product defects) into the influence maximization

problem.

METHODOLOGY :

Given (a) a labeled social network G= (V, E, L), in which V is the vertex set, E is the edge set, L is the set of labels L= {l_1, l_2, ……, l_m}, and each vertex iεV is associated with a set of labels LB(i), which is a subset of L, (b) a budget k (integer) representing the number of selected seed nodes, (c) a set of targeted labeled L_t, which is a subset of L, representing the information of targeted consumers, and (d) a designated profit value pf_{la} for each targeted label l_a ε L_t, the labeled influence maximization problem is to find a set of seed nodes S in V,|S|=k such that the labeled influence spread (i.e., total profit) caused by S, denoted by $R_{Lt}(S)$, is maximized, where $R_{Lt}(S)= \sum_{i\varepsilon Q} pf_{LB(i)}$, Q={i | iε A(S), LB(i) ε L_t} and A(S) is the set of nodes activated by the seed setS.

The General Greedy can be extended to solve the labeled influence maximization problem. We consider the target label L_t when estimating the labeled influence spread of certain seed set. And thus the expected number of activated nodes R(S) is replaced by RLt(S). Then the greedy method is the same to select the node $v_t$which maximizes of the marginal gain of labeled influence spread, v_t= $argmax_{v\varepsilon V\setminus St-1}${$R_{Lt}$ (St-1 U {v}) - R_{Lt} (St-1}. In addition, based on Kempe et al.'s study we can further show R_{Lt} is a submodular function:

$$\mathcal{R}_{L_t}(S \cup \{v\}) - \mathcal{R}_{L_t}(S) \geq \mathcal{R}_{L_t}(T \cup \{v\}) - \mathcal{R}_{L_t}(T),$$

where the seed set S, is a subset of T and v ε V.

The Labeled General Greedy algorithm is described in Algorithm 1.

Algorithm 1. Labeled General Greedy (G, k)

```
Initialize S = ∅ and Round = 10000.
for t = 1 to k do
    for each node v ∈ V\S do
        s_v = 0.
        for i = 1 to Round do
            s_v += R_{L_t}(S ∪ {v}).
        s_v = s_v/Round.
    S = S ∪ {argmax_{v∈V\S}{s_v}}.
Return S.
```

The New Greedy algorithm is proposed by Chen et al. to improve the efficiency of the general greedy algorithm for original influence maximization problem. They exploit the assumption of the independent cascade model: each active node u has only one chance to activate one of its inactive neighbor v, and no matter u successfully activate v or not, u will never make an attempt to activate v again. Due to all edges has only one chance to be propagated, their method determines which edges will be propagated in prior according to the probabilities on edges. And thus those nodes reachable from one to another are considered to be able to influence each other. Such strategy significantly reduces the execution time spent on the simulation of influence spread.

The proposed Labeled New Greedy algorithm is described in Algorithm 2. First, based on the influence probability p on each edge, we determine which edges are selected in prior. Those edges do not be selected are removed from the graph, and a trimmed network G' is derived. We use $T_{G'}(S)$ to represent the node set which can be successfully activated by the seed set S in G'. And those nodes reachable from the seed set S in G' are those ones can be successfully influenced. Using the depth-first search, we can easily derive $T_{G'}(S)$ $T_{G'}(v)$. For each v ε V\S, we can obtain the set $T_{G'}(v)$ of nodes which can be activated by nodev. And the labeledinfluencespreadcanbederivedby|$T_{G'}(v)$| = $\sum_{u\varepsilon TG'andLB(u)\varepsilon Lt} pf_{LB(u)}$.

Algorithm 2. Labeled New Greedy (G, k)

```
Initialize S = ∅ and Round = 10000.
for t = 1 to k do
    Set s_v = 0, for all v ∈ V\S.
    for i = 1 to Round do
        Derive G' by removing each edge from G according to
        the probability 1 − p.
        Compute Γ_{G'}(S).
        Compute |Γ_{G'}(v)|, for all v ∈ V.
        for each node v ∈ V\S do
            if v ∉ Γ_{G'}(S) then s_v = |Γ_{G'}(v)|.
    Set s_v = s_v/Round, for all v ∈ V\S.
    S = S ∪ {argmax_{v∈V\S}{s_v}}.
Return S.
```

Other than the New Greedy algorithm, Chen et al. [1] further propose the Degree Discount heuristic to efficiently find the effective seed

nodes. Degree Discount assumes that the propagation of influence has lower potentials to spread globally. And thus it is natural to consider only one-step neighbor nodes and select nodes with high degree values, which tends to has higher expectations of influence, to be the seed ones. The central idea is to compute and update such expectations of influence in each round of selection. If one wants to select k seed nodes, Degree Discount will be performed k times. After selecting node w as a seed in each round, Degree Discount will recalculate theexpectationofinfluenceofeachw'sneighborvbec ausev'sexpectationofinfluence will get discount due to selecting w as the seed. The formula for the recalculation of the expectation of influence E_V is E_V = (1-p) |tv| x (1+ (|d_V| - |t_V|) x p), where d_V = {u | u ε V & (v, u) ε E}, |d_V| is the number of neighbors of v, t_V = {u | u ε V & (v, u) ε E & u ε S}, |t_V| is the number of nodes which belongs to v's neighbors and are contained by the seed set S. In other words, E_V is the expectation value that v is not only never influenced by existing seed nodes, but also able to activate those do not be selected as the seed nodes yet.

To address the labeled influence maximization problem, we modify the Degree Discount heuristic by considering the label information. Algorithm 3 describes the proposed Labeled Degree Discount algorithm.

Algorithm 3. Labeled Degree Discount (G, k)

Initialize $S = \emptyset$.
for each node $v \in V$ **do**
  Compute its degree d_v
  $dd_v = d_v$.
  Initialize $|t_v| = 0, |r_v| = 0$.
for $t = 1$ to k **do**
  Select $u = argmax_{v \in V \setminus S}\{dd_v\}$.
  $S = S \cup \{u\}$.
  for each $v \in N(u)$ and $v \in V \setminus S$ **do**
    if $LB(u) \in L_t$ **then**
      $|t_v| = |t_v| + 1$.
    else do $|r_v| = |r_v| + 1$.
    $dd_v = (1-p)^{|t_v|+|r_v|} \cdot (1 + \sum_{u \in d_v - t_v}(pf_{LB(u)} \cdot p))$.
Return S.

Maximum Coverage Greedy

In real-life cases, the social network can be obtained before marketers employ it to plan advertising strategies. Hence, for the original influence maximization problem, we can perform some preprocessing to find the top-k influencers in advance so that marketers can determine the advertising plans without waiting for the execution. The proposed Maximum Coverage method allows offline computing influence potential between nodes and online finding the top-k seed nodes for target labels and corresponding profits.

The basic idea of our Maximum Coverage method is to consider the topological proximities between nodes as the potentials of successful influence from one to the other. If node u has higher proximity towards node v, we think u will have a higher probability to activate v. For the computation of proximity scores between nodes in a network, there are many proposals for different scenarios. To integrate with the influence propagation in the independent cascade model, in this work, we modify the Cycle Free Effective Conductance (CFEC) method to compute the pairwise proximity scores. The CFEC method applies the random walk mechanism to compute the probability that one node surfs in the graph and ever arrives at the other.

Based on the offline computed pairwise proximity scores, we propose another greedy method to solve the labeled influence maximization problem. The central idea aims to consider the original problem as the Maximum Coverage Problem. The desired set of seeds, consequently, will be the set of nodes that can cover nodes with maximum total profit.

Given a graph G = (V, E), for each viε V, it has a set of nodes U_i = {v1, v2, ..., vm} is a subset of V, where the proximity score P_{cf}(vi, vj) ≠ 0 (vj ε U_i, vj ≠ vi) . Let U be the set of $U_i$for each viε V, i.e., U = {U1, U2, ..., U_n} and $U_i$is a subset of V . The original maximum k coverage problem is to find a set U' which is a subset of U and |U'| = k such that the coverage (i.e., number) of U',

which equals to |UUl ε U, Ul|, is maximized. By applying such concept to the labeled influence maximization problem, the goal aims to find a seed set U' which is a subset of U and |U'| = k such that the labeled influence spread (i.e., total profit) of U', which equals to $\sum_{U_l \in U'} \sum_{v_j \in U_l} pf_{LB(v_j)}$, is maximized.

Experiments

We conduct the experiments to evaluate the effectiveness and efficiency of the methods for the proposed labeled influence maximization problem. We employ the Internet Movie Database (IMDb) to construct the labeled social network. Actors and actresses of movies during

We demonstrate the results when the online queries contain multiple target labels with equal profit weights. Figure 2 shows the effectiveness of the query consisting of one target label {drama}. The results exhibit as the seed number further increases, the Maximum Coverage is generally the better one and competitive to the Labeled New Greedy method. However, the Labeled New Greedy suffers from the time efficiency, as will be shown in Table 1.

We designate the online queries contains multiple target labels with different profit weights. Figure 3 shows the effectiveness of that the targeted labels are drama and comedy with profit weights 1 and 3 respectively. In general, the Labeled New Greedy outperforms the other three ones. We can find, again, the Labeled New Greedy is the best one, though its execution time is infeasible for online marketing analysis (as shown in Table 1). The more feasible is the Maximum Coverage method, which can not only make higher total profits but also allow real-time querying for online marketingtasks.

Table 1 presents the execution time (in second) for both the three methods. The results are derived under the case that the targeted label contains only comedy with single profit weight since the other cases show the same trends. We can find that Labeled Degree Discount is fastest one, which even needs no more than one second, because it considers only the one-step neighbors. The time efficiency of the Maximum Coverage is acceptable. It also has promising effectiveness in different cases, as shown in previous elaborations. Though the Labeled New Greedy performs well in some cases, its run time is totally infeasible for online marketing analysis.

CONCLUSION :

We introduce the labeled influence maximization problem in social networks for target marketing which focuses on target customers. We propose three algorithms to find the seeds for maximizing the influence spread. The experiment result suggests that the proposed methods perform faster while maintaining the influence spread guarantee. Future work includes the development of algorithms under weighted cascade model and linear threshold models.

REFERENCES :

W.Chen,Y.Wang,andS.Yang,"EfficientInfluenceMaximizationinSocialNetworks,"Proc. of ACM International Conference on Knowledge Discovery and Data Ming KDD,2009.

D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 137–146,2003.

Joshua Rubin Abrams, Jose Celaya-Alcala, Drew Baldwin Ryan Gonda, Zhaoren Chen "Analysis of Equity Markets: A Graph Theory Approach"

Chen W, Castillo C, Lakshmanan LV (2013) Information and Influence Propagation in Social Networks (Morgan & Claypool Publishers, San Rafael, CA).

Table 1. The time efficiency of $L_t = \{Comedy\}, pf_{Comedy} = 1$.

	Labeled NewGreedy	Labeled Degree Discount	Maximum Coverage
Running Time	8 hours	0.125 sec.	20.3 sec.

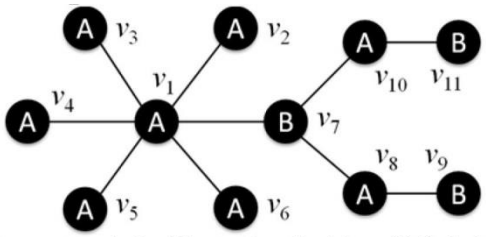

Figure 1. A toy network for illustrating the idea of labeled influence maximization.

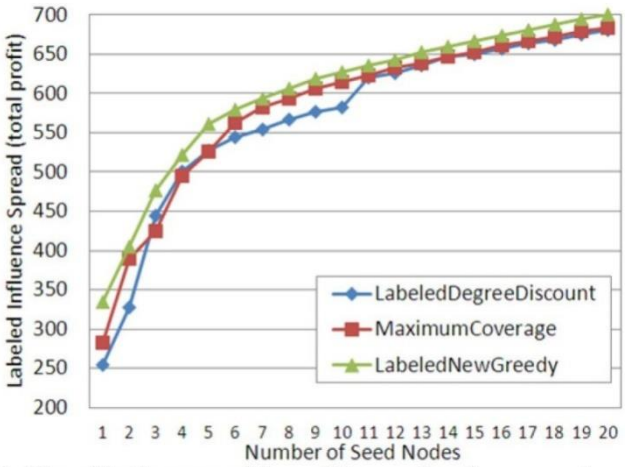

Figure 2. The effectiveness of $L_t = \{Drama\}, pf_{Drama} = 1$.

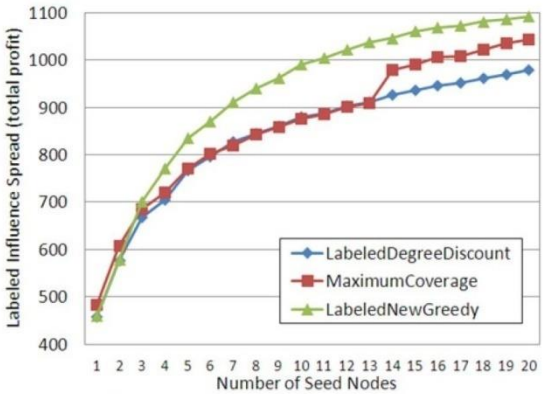

Figure 3. The effectiveness of $L_t = \{Drama, Comedy\}, pf_{Drama} = 1, pf_{Comedy} = 3$.