The Research and Design of Algorithm for Influence Maximization in Social Networks Based on Activity*

Lianying ZHOU, Yuting PAN*, Xin ZHENG, Yadong JIN

School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China

Abstract

Influence maximization problem is how to select a set of seed nodes and then maximize the dissemination of information according to some strategies in social networks. In this paper, we introduce the node activity into the IC model and then propose the AIC (activity independent cascade) model and the corresponding heuristic algorithm ACH which narrows the range of primaries and improves the quality of candidate nodes by considering node activity attributes. Our simulation results also show its superiority-ACH is close to KK on influence scope while better than other heuristic algorithms on timeliness, which manifests the applicability of AIC for large social networks.

Keywords: Influence Maximization; Activity; Social Network; Heuristics

1 Introduction

Influence maximization is to disseminate and diffuse information as widely as possible, which has been summarized as an algorithm problem [1] by Domingos and Richardson. And it is converted to be a problem of how to select K initial nodes, by which we can maximize the spreading range in networks [2].

To solve the problem, Kemple and Kleinberg propose a greedy algorithm [3] which selects the most influential nodes as initial ones in every step to guarantee the optimal spreading within (1-1/e) at the expense of timeliness. By comparison, heuristic algorithms improve the influence scope on the basis of its timeliness, which proves to be more suitable for large-scale social networks.

Information dissemination follows certain rules or called models. Nowadays, the most widely used ones are the Independent Cascade (IC) and the Linear Threshold (LT) [3]. Recently, the research work of heuristics mainly focuses on the sub-modularity [4] under the IC model. It improves the efficiency of the greedy algorithm by reducing its range increment. Ideally, activated nodes in traditional IC models are considered 100% to activate its neighbor nodes. In fact, the activity attribute of nodes has a significant impact on node dissemination. Thus, we propose an
Activity Independent Cascade model introducing the node activity into the IC and the corresponding algorithm ACH. The algorithm is demonstrated to be close to KK on influence scope while better than other heuristic algorithms on timeliness.

This paper is organized as follows. The next section gives the background of the problem which includes the introduction of classic information dissemination models and algorithms; Section 3 describes the AIC model and the corresponding heuristic algorithm ACH; Section 4 demonstrates the superiority of the ACH algorithm based on the analysis of our experimental result; Section 5 contains some conclusions plus some ideas for further work.

2 Backgrounds

2.1 Tow classic models

Nowadays, the influence maximization problem in social networks has been a hotpot. To solve the problem, there are some classic influence propagation models. Followings are the description of Independent Cascade (IC) model and the Linear Threshold (LT) model.

LT [5, 6] is a cumulative model, in which every node has an activation threshold value. Once the sum of the influence of its neighbors is larger than the value, the node is successfully activated. Influence in the model can be accumulated if the node does not be activated till it is activated or the spreading process ends. But the cumulative attribute increases the calculation and complicates the node network.

IC model [7, 8] is a probabilistic model, in which whether a node can be activated is a probabilistic event. Different from LT, nodes try to activate its neighbor nodes only once no matter it is successfully activated or not. In this case, the amount of calculation can be reduced comparing with LT. In addition, the activation is random and independent and the order also has nothing to doing with the results.

2.2 Tow types of classic algorithms

On the basis of the influence maximization model, a variety of algorithms have been proposed to find out the most influential seed nodes. There are generally two types of algorithms: greedy algorithms and heuristics.

The greedy algorithm [3] proposed by Kemple and Kleinberg calculates the marginal influence of all inactive nodes in every step to guarantee the optimal influence spreading within (1-1/e). However the calculation complexity increases dramatically along with the expansion of online social network scales.

Due to the serious drawback on efficiency in the greedy algorithm, some scholars recommend to focus our efforts on improving the influence spreading of heuristic algorithms. Wei Chen proposes the Degree Discount IC algorithm [9]. The general idea is as follows. When considering \(v\) as a new seed node, we should discount the degree of \(v\) its neighbor node \(u\) has been in the seed set. So when a node is selected as seed node, we should update the out-degree of neighbor nodes of the node.

\[
dd_v = d_v - 2t_v - (d_v - t_v)t_v p, \quad v \in N(u), \quad v \in V \setminus S
\]  

(1)
\( N(u) \) is the set of Neighbor nodes of node \( u \), \( dd_v \) is the discounted out-degree of node \( v \), \( d_v \) is the out-degree of node \( v \), \( t_v \) is the number of neighbors of node \( v \) that are already selected as seeds, \( p \) is the probability which the node \( v \) is activated by the node \( u \). Compared with other traditional heuristic algorithms, Degree Discount IC algorithm achieves a great breakthrough in the influence spread.

3 AIC Model and ACH Algorithm

3.1 The main idea of the AIC model

Generally, the influence of a node in IC model is determined by its out-degree and influence probability and an activated node is considered 100% to activate its neighbor nodes. Actually, some influential nodes are inactive and do not have a good performance in the dissemination and diffusion of influence. To address such issue, we propose the Activity Independent Cascade model, in which we take into consideration the activity of nodes during the spreading process. Thus we can select both influential and active nodes.

In AIC model, each edge \((u, v) \in E(u, v \in V)\), which is associated with an influence probability \( p(u, v) \), representing the probability of node \( v \) activated by the activated node \( u \). The process of the dissemination and diffusion of influence under AIC model is similar to the IC model: \( A_t \subseteq V \) is the set of nodes activated at every step \( t \geq 0 \), with \( A_0 = S \). In step \( t + 1 \), each node \( u \in A_t \) only has one chance to activate all of its currently inactivated out-neighbors \( v \) with probability \( p(u, v) \). The process continues until \( A_t = \varnothing \). The difference between the two models is that the activated node with ACT \((ACT \in [0, 1])\) executes the behavior of activating its neighbors-out according to the ACT of the node. The higher the value of the ACT, the higher the probability is.

3.2 Implementation of ACH algorithm

In this paper, we propose a new heuristic algorithm called ACH based on the AIC model to solve the problem of influence maximization considering the ACT of the node.

The main idea of the algorithm is to simplify the size of social network first and then reduce the amount of calculation. There are mass of nodes and links between nodes in a social network. In fact, there are only a few ones having the chance to be selected as seed nodes. If the probability value that the node \( v \) is activated by the activated node \( u \) is too low, we can ignore the link between them. That is to say, there is no need to calculate the influence of every node in the social network.

We can prune the network when build it through filtering out the nodes and the links which are not helpful for selecting seed nodes. It greatly reduces the size of the network, thereby reducing the consumption of computing. Then in the streamlined network \( G' \), considering the ACT of nodes, we select the nodes with high value of ACT into the set of candidate nodes named H. Then we calculate the integrated influence of every node in \( H \), which is named as \( \text{aps}(v) \) based on the ACT of nodes and the influence of nodes. The node with high value of \( \text{aps}(v) \) can be selected as seed nodes.

Of course, when calculating the \( \text{aps}(v) \) of node \( v \), we also need to consider the ACT of out-
neighbors of \( v \). As a node is selected as seed node, we simulate the process of the influence propagation and labeled each edge in the process. We can ignore these edges and update the out-degree of the related nodes, because each node has single chance to activate any of its currently inactivate out-neighbors, so that we can improve the accuracy of selection. The steps of selection are as follows.

(1) Build the streamlined network

\[
G' = (V', E')
\]

\[
V' = \{v|v \in V, d_v > \theta\}, E' = \{(u, v) \in E|p(u, v) > \beta\}
\]

(2) Build the set \( H \)

\[
H = \{v \in V'|ACT_v \geq \gamma\}
\]

(3) Calculate the \( \text{aps}(v) \) of nodes in \( H \)

\[
\text{aps}(v) = ACT_v \ast \sum_{u \in N(v)} d_u \ast p(v, u), v \in H
\]

(4) Select the node with high value of \( \text{aps}(v) \) as seed nodes

\[
S = S \cup \{u\}, u = \arg \max_v \{\text{aps}(v)|v \in H \setminus S\}
\]

(5) Simulate the process of the influence propagation and labeled each edge in the process

\[
\text{flag}(u, v) = \text{true}, (u, v) \in E'
\]

(6) Update the nodes and the links

\[
G' = (V'', E'')
\]

\[
V'' = \{u \in V'|u \notin A\}, E'' = \{(u, v) \in E'|\text{flag}(u, v) = \text{false}\}
\]

(7) Update the out-degree

\[
d'_v = d_v - t_v, t_v = \sum_{\text{flag}(v, u) = \text{true}, u \in N(v)} 1
\]

(8) Update the \( \text{aps}(v) \) of nodes

\[
\text{aps}(v) = ACT_v \ast \sum_{u \in N(v)} d'_u \ast p(v, u)
\]

The ACH algorithm is described as follows:

## 4 Experiments and Evaluations

We conduct extensive experiments on several real-world and synthetic networks, comparing the four algorithms on quality of seeds sets and running time. Through the experimental results and analysis, we demonstrate that our heuristic algorithm ACH matches the KK greedy algorithm in influence spreading and has a better performance in efficiency than other heuristics.
Algorithm ACH: \[ACH(G = (V, E), k, ACT, \theta, \beta, \gamma)\]

1: initialize \(S = \varnothing\)
2: initialize \(G' = (V', E'), V' = \{v|v \in V, d_v > \theta\}, E' = \{(u, v) \in E|p(u, v) > \beta\}\)
3: \(H = \{v \in V'|ACT_v \geq \gamma\}\)
4: for each \(v \in H\) do
5: initialize \(aps(v) = ACT_v \times \sum_{u \in N(v)} d_v \times p(v, u), v \in H, ACT_u > \gamma\)
6: for \(i = 1\) to \(k\) do
7: select \(u = \arg \max_v \{aps(v)|v \in H\setminus S\}\)
8: \(S = S \cup \{u\}\)
9: for each \((u, v) \in E'\) do
10: \(\text{flag}(u, v) = \text{true}\)
11: \(G' = (V'', E'')\)
12: \(V'' = \{u \in V'|u \notin A\}\)
13: \(E'' = \{(u, v) \in E'|\text{flag}(u, v) \neq \text{true}\}\)
14: for each neighbor \(v\) of \(u\) and \(v\) and \(S\setminus V\) do
15: \(t_v = \sum_{\text{flag}(v, u) = \text{true}, u \in N(v)} 1\)
16: \(d'_v = d_v - t_v\)
17: \(aps(v) = ACT_v \times \sum_{u \in N(v)} d_v \times p(v, u)\)
18: end for
19: end for
20: output \(S\)

4.1 Experimental preparation

We select two large real-world networks, Epinions and DBLP, as the data sets for our experiments. The statistics of the datasets are presented in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Epinions</th>
<th>DBLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of nodes</td>
<td>75K</td>
<td>655K</td>
</tr>
<tr>
<td>Num. of edges</td>
<td>655K</td>
<td>2.0M</td>
</tr>
<tr>
<td>Average degree</td>
<td>13.4</td>
<td>6.1</td>
</tr>
<tr>
<td>Maximum degree</td>
<td>3079</td>
<td>588</td>
</tr>
</tbody>
</table>

We compare the four algorithms on quality of seeds sets and running time based on these two data sets, including Greedy, Degree, MIA and ACH algorithm. We compare ACH with Greedy in influence spreading and other two heuristics in running time. We generate a random number
between 0 and 1 as the value of the ACT of a node under AIC model, based on the out-degree of
the node. In 80% of the case, the value of the ACT of a node is proportional to the out-degree of
the node, and in 20% of the case, the value of the ACT of a node is inversely proportional to the
out-degree. The average out-degree is defined as \( \theta \), the average activation probability is defined
as \( \beta \) and the average value of ACT is defined as \( \gamma \).

4.2 Experimental results and analysis

In this paper, we compare the four algorithms on influence spreading and efficiency by comparing
the quality of seeds sets and the running time. Greedy is so time-consuming that it finishes on
the two data sets with 3 days. Thus, we just compare its influence spreading with the three
heuristics.

The quality of seeds sets is evaluated through the expected number of the activated nodes. We
initialize the network \( G = (V, E) \) by generating the value of ACT of nodes according to certain
rules under AIC model. Then we find the seeds sets \( S \) in \( G = (V, E) \) with the four algorithms.
Fig. 1 illustrates that according to the influence spreading achieved on Epinions, ACH has the
highest seeds set quality, whose influence spreading is very close to Greedy. As shown in Fig. 2,
we obtain a similar result on DBLP. Fig. 3 illustrates that when we define \( ACT = 1 \), ACH still
has good performance. So we can see that ACH has good stability.

![Fig. 1: Influence spread on Epinions](image1)

![Fig. 2: Influence spread on DBLP](image2)

We judge the efficiency of an algorithm by the running time of selecting the seeds set. We
demonstrate the running time of the four algorithms results on the two datasets in Table 2. The
Greedy spends too much time and the Degree finishes almost instantly in all test cases. So they
are not included in the table. We simplify the social network by filtering out nodes and links
which are inefficient on seed nodes selecting. After that, we can see that ACH is more efficient
than MIA because its running time is just about half of MIA.

5 Conclusions

In this paper, we introduce the node activity into the IC model and then propose the AIC model
and the corresponding heuristic algorithm ACH to improve influence spread, which is demonstrat-
Fig. 3: Influence spread on DBLP when $ACT = 1$

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Epinions</th>
<th>DBLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Degree</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MIA</td>
<td>15s</td>
<td>41s</td>
</tr>
<tr>
<td>ACH</td>
<td>7s</td>
<td>23s</td>
</tr>
</tbody>
</table>

