Negative Influence Minimizing by Blocking Nodes in Social Networks

Senzhang Wang¹, Xiaojian Zhao¹, Yan Chen¹, Zhoujun Li¹, Kai Zhang¹, Jiali Xia²
¹State Key Laboratory of Software Development Environment
Beihang University, Beijing 100191, China
²School of Software and Communication Engineering
Jiangxi University of Finance and Economics, Nanchang 330013, China
{szwang@cse, lizj@cse, chenyan@cse}@buaa.edu.cn, zhaoxj01@gmail.com,
zhangkai.plus@foxmail.com, xiajl65824@263.net

Abstract

Social networks are becoming vital platforms for the spread of positive information such as innovations and negative information propagation like malicious rumors. In this paper, we address the problem of minimizing the influence of negative information. When negative information such as a rumor emerges in the social network and part of users have already adopted it, our goal is to minimize the size of ultimately contaminated users by discovering and blocking uninfected users. A greedy method for efficiently finding a good approximation solution to this problem is proposed. The comparison experimental results on the Enron email network dataset demonstrate our proposed method is more effective than centrality based methods, such as degree centrality, betweenness centrality and PageRank.

Introduction

Online social networks are becoming vital platforms for information propagation. It not only includes positive information such as innovations and hot topics (Chen et al. 2012), but also negative information like malicious rumors and disinformation. The research on maximizing the influence of positive information, namely Influence Maximization Problem, has attracted remarkable attention recently due to its novel idea of leveraging some social network users to propagate the awareness of products (Kempe, Kleinberg, and Tardos 2003; Chen, Wang, and Wang 2010; Shirazipourazad et al. 2012). However, the problem of reducing the influence of negative information or undesirable things gets less attention, although it is an important research issue. Take the rumor for example, even with a small number of its initial adopters, the quantity of the ultimately infected users can be large due to triggering a “word-of-mouth” cascade in the network. Therefore, it is an urgent research issue to design effective strategies for reducing the influence coverage of the negative information and minimizing the spread of the undesirable things.

Some related research work has been made on minimizing the influence of negative information. Baumgartner et al. proposed a links blocking method to minimize the expected contamination area of the network (Baumgartner, Gottlob, and Flesca 2008). However, the fact of some nodes infected is not considered. Habiba et al. addressed the problem of finding spread blockers in dynamic networks, and they found good blockers are simply those nodes with high degree (Habiba et al. 2010). Budak et al. investigated the problem of influence limitation where a “bad” campaign starts propagation from a certain node in the network and use the notion of limiting campaigns to counteract the effect of misinformation (Budak, Agrawal, and Abbadi 2011). Different from previous work, our research concerns more about a specific contamination scenario in the network, and how to minimize the negative influence by blocking a small set of uninfected nodes.

We formally define the Negative Influence Minimization Problem as follows. Given a network represented as a directed graph $G = (V, E)$, we define $V$ and $E \subseteq V \times V$ as the sets of nodes and edges in $G$, respectively. Assume negative information spreads in the network and some nodes $I \subseteq V$ are infected, our goal here is minimizing the number of ultimately infected nodes by blocking $k$ uninfected nodes $S \subseteq \{V \setminus I\}$, where $k$ is a given const. It can be represented as the following optimization problem:

$$\begin{align*}
\text{Minimize} & \quad \sigma\{I\setminus S\} \\
\text{subject to} & \quad |S| \leq k.
\end{align*}$$

where $\sigma\{I\setminus S\}$ denotes the influence (number of ultimately infected nodes) of $I$ when the node set $S$ is blocked.

Methodology

In this section, we propose a greedy algorithm based on maximum marginal gain rule for this problem. To demonstrate the effectiveness of it, we compare it against three classical centrality based influence evaluation methods.

As the basis for all methods, Independent Cascade (IC) Model is adopted as the information diffusion model in this paper. In IC model, each edge is given a probability value $p_{u,v}(\leq 1)$, which corresponds to the probability that node $u$ influences node $v$. The diffusion proceeds from a given initial active set $S_0$ in the following way: at each diffusion step $t$, each node which first becomes active at the step $t - 1$ has exactly one chance of activating each of its inactive neighbors according to edge probabilities. The propagation process will be finished when no new nodes become active.
Proposed Method. The proposed method is described as Algorithm 1. We first start with the empty set $S_0 = \emptyset$, then add node $x_i$ with the maximum marginal gain in the step $i$, until $k$ nodes are selected. Here, we define the maximum marginal gain rule as the node selection rule which maximizes the decrement of ultimately infected nodes number when the selected node is added to the blocking node set $S$. We formally represent it as follow:

$$x_i = \arg \max_{x \in \{ V \backslash S_{t-1} \}} \sigma \{ I \backslash \{ S_{t-1} \} \} - \sigma \{ I \backslash \{ S_{t} \} \}$$

(2)

Algorithm 1. ConMinGreedy$(G = (V, E), I, k)$

1: initial $S_0 = \emptyset, \Delta_v = 0$
2: for $i=1$ to $k$ do
3: for each vertex $v \in \{ V \backslash \{ S_0 \} \}$ do
4: $\Delta_v = \sigma \{ I \backslash \{ S_{i-1} \} \} - \sigma \{ I \backslash \{ S_{i} \} \}$
5: end for
6: $S_i = S_{i-1} \cup \{ v_* : \arg \max_{v \in \{ V \backslash \{ S_0 \} \}} \Delta_v \}$
7: end for
8: output $S_k$

Three classical centrality based methods are used as baselines and described as follows:

- **Out-Degree.** The out-degree $d(v_i)$ of a node $v_i$ is the number of outgoing links from the node $v_i$, Kempe et al. showed high degree nodes may outperform other centrality-based heuristics in terms of influential identification(Kempe, Kleinberg, and Tardos 2003).

- **Betweenness Centrality.** A node’s betweenness is equal to the number of shortest paths from all nodes to all others that pass through that node. Recently, Betweenness centrality has become an important centrality measure in social network(Brandes 2008).

- **PageRank.** PageRank method is a representative eigenvector centrality method. For its excellent performance in the web pages ranking, we also use it as a baseline(Page et al. 1999).

In the comparison experiment, we select and block top-$k$ uninfected users according to their out-degree, betweenness and PageRank authority, respectively.

Experimental Results

Utilizing the real email communication network, we experimentally evaluate the performance of our proposed approach. Enron email communication network covers all the email communication within a dataset of around half million emails from Enron corpus. Nodes of the network are email addresses and if an address $i$ sent at least one email to address $j$, the graph contains an undirected edge from $i$ to $j$. This network contains 36,692 nodes and 367,662 edges.

In the IC Model, we assign a uniform probability to each edge of the graph. Two propagation probabilities, $p = 0.2$ and $p = 0.4$, are used in our experiment. To further investigate the effectiveness of our method, we also conduct experiments on different number of initial contaminated nodes $|I|$. In addition, besides the three centrality based methods, random selection method is also used as comparison. Figure 1 shows the experimental results by varying the number of blocked nodes. Parameters are set to $p = 0.2$ $|I| = 100$, $p = 0.2$ $|I| = 200$, $p = 0.4$ $|I| = 100$ and $p = 0.4$ $|I| = 200$, respectively. The results show that the proposed approach outperforms all the other methods on all the four groups of experiment. The performance of PageRank is slightly better than that of betweenness centrality and degree centrality. However, it is inferior to our proposed method. It is probably that these methods ignore the initial infected nodes, which result in ineffective in some scenario, especially when the selected nodes are far from the nodes in $I$.

Conclusions

In this paper we investigated the problem of minimizing the spread of negative information by blocking nodes in social networks and proposed a greedy method for efficiently finding a good approximate solution. Comparison with three classical centrality based baseline methods, our method achieved a significant improvement on a real email communication network dataset. Ongoing work focuses on (a) the theory guarantee of the method (whether it is a matroid) and (b) how to extend it to a dynamic network where the number of infected nodes increases with time.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant No. 61170189, 61202239], and the Fund of the State Key Laboratory of Software Development Environment [Grant No. SKLSDE-2013ZX-19].

1http://snap.stanford.edu/data/email-Enron.html
References


