Topic-aware Social Influence Minimization

Qipeng Yao[1,2], Chuan Zhou[2], Ruisheng Shi[1], Peng Wang[2]and Li Guo[2]
[1]Education Ministry Key Laboratory of Trustworthy Distributed Computing and Service,
Beijing University of Posts and Telecommunications, Beijing, China
[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
yaoqipeng0706@gmail.com,
{zhouchuan,guoli,wangpeng}@iie.ac.cn,shiruisheng@bupt.edu.cn

ABSTRACT

In this paper, we address the problem of minimizing the negative influence of undesirable things in a network by blocking a limited number of nodes from a topic modeling perspective. When undesirable thing such as a rumor or an infection emerges in a social network and part of users have already been infected, our goal is to minimize the size of ultimately infected users by blocking k nodes outside the infected set. We first employ the HDP-LDA and KL divergence to analysis the influence and relevance from a topic modeling perspective. Then two topic-aware heuristics based on betweenness and out-degree for finding approximate solutions to this problem are proposed. Using two real networks, we demonstrate experimentally the high performance of the proposed models and learning schemes.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining

General Terms

Method, Performance

Keywords

Influence Minimization; Blocking Nodes; Social Networks

1. INTRODUCTION

The online social networks can launch diffusion including not only positive information such as innovations, hot topics, and novel ideas, but also negative information like malicious rumors and disinformation [2]. 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.

This problem has received a good deal of attention by the data mining research community in the last decade [3, 4], but quite surprisingly, the characteristics of the item being the subject of the influence minimization has been left out of the picture.

In this paper, we aim to minimize the spread of an existing undesirable thing by blocking a limited number of nodes in a network from a topic modeling perspective. More specifically, when some undesirable thing starts with some initial nodes and diffuses through the network under the topic-aware independent cascade (TIC) model, we consider finding a set of k nodes such that the resulting network by blocking those nodes can minimize the expected contamination area of the undesirable thing, where k is a given positive integer. We refer to this combinatorial optimization problem as the *influence minimization problem*. For this problem, we fisrt employ the HDP-LDA and KL divergence to analysis the authoritativeness, influence and relevance from a topic modeling perspective. Then we propose two topic-aware heuristics based on betweenness and out-degree for finding approximate solutions to the problem. With two large real networks including Sina microblog and Facebook, we experimentally demonstrate that the proposed topic-aware node-removal heuristics outperform the well-studied notions of centrality measures.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742767.

2. PROBLEM FORMULATION

To model the topic-aware social influence, we adopt the *Topic-aware Independent Cascade (TIC) Model* [1], where the user-to-user influence probabilities depend on the topic. Therefore, for each arc $(v,u) \in E$ and each topic $z \in [1,K]$ we are given a probability $p^z_{v,u}$, representing the strength of the influence exerted by user v on user u on topic z. Moreover for each item $i \in \mathcal{I}$ that propagates in the network, we have a distribution over the topics, that is for each topic $z \in [1,K]$ we are given $\gamma^z_i = P(Z=z|i)$, with $\sum_{z=1}^{K} \gamma^z_i = 1$. In this model a propagation happens like in the IC model: when a node v first becomes active on item i, has one chance of influencing each inactive neighbor u, independently of the history thus far. The tentative succeeds with a probability that is the weighted average of the link probability w.r.t. the topic distribution of the item i: $p^i_{v,u} = \sum_{z=1}^{K} \gamma^z_i p^z_{v,u}$.

Under the directed graph $G=(V,E)$, the *influence spread* of the initially infected set S, which is the ultimately expected number of infected nodes, is denoted as $\sigma(S|V)$.

Now we present a mathematical definition for the *influence minimization problem*. Assume the negative information spreads in the network $G=(V,E)$ with initially infected nodes $S \subseteq V$, our goal here is to minimize the number of ultimately infected nodes by blocking k nodes (or vertices) of set $D \in V$, where k ($\ll |V|$) is a given const. It can be formulated as the following optimization problem:

$$D^* = \arg \min_{D \subseteq V, |D| \leq k} \sigma\big(S\big|V \backslash D\big) \tag{1}$$

where $\sigma(S|V\backslash D)$ denotes the influence (number of ultimately infected nodes) of S when the node set D is blocked.

3. ANALYSIS AND SOLUTION

The problem of learning the parameters of the TIC models takes in input the social graph $G = (V, E)$, a log of past propagations $\mathbb{D}$, and an integer K, which can be learnt by the Latent Dirichlet Allocation based on Hierarchical Dirichlet Process (HDP-LDA) method. The propagation log is a relation (`User,Item,Time`) where a tuple $(u, i, t) \in \mathbb{D}$ indicates that user u adopted item i at time t. The output of the learning problem is the set of all parameters of the TIC propagation model, which we denote Θ: these are γ_i^z and $p_{v,u}^z$ for all $i \in \mathcal{I}$, $(v, u) \in E$, and $z \in [1, K]$. Assuming that each propagation trace is independent from the others, the likelihood of the data given the model parameters Θ, can be expressed as: $\mathcal{L}(\Theta; D) = \sum_{i \in \mathcal{I}} \log \mathcal{L}(\Theta; D_i)$. We then adopt the standard EM inference of parameters Θ for TIC. We calculate the topic distributions of each uninfected node w and negative information i via HDP-LDA, then calculate the KL divergences $d(w, i)$ between node w and information i from the topic perspective.

Now we are back to the optimal problem (1), any straightforward method for exact solution suffers from combinatorial explosion for a large network. Therefore, we consider approximately solving the problem, while a natural idea is to block the nodes in the neighborhood of infected set. Specifically, given the initially infected set S and the negative information i, define the neighborhood set $N(S)$ like

$$N(S) := \{v \in V \backslash S : \exists u \in S, \ s.\ t.\ (u, v) \in E\}.$$

We want to block k susceptible nodes in the set $N(S)$ to minimize the negative influence. Since the set $N(S)$ is usually very large (i.e. $|N(S)| \gg k$), a natural question arises, *how to select k susceptible nodes from the set $N(S)$ to block in order to make the ultimate influence as small as possible*? In this paper, given the negative information $i \in \mathcal{I}$, we introduce two scoring methods for the nodes in $N(S)$, and then select k nodes with the highest scores as the objectives to block.

Topic-aware Betweenness scoring method. Given the initially infected nodes S, the betweenness score $b(w)$ of a node $w \in N(S)$ is defined as follows:

$$b(w) = \sum_{u \in S, v \in V \backslash S} \frac{n(w; u, v)}{N(u, v)} \tag{2}$$

where $N(u, v)$ denotes the number of the shortest paths from node u to node v in G, and $n(w; u, v)$ denotes the number of those paths that pass w. Here we set $n(w; u, v)/N(u, v) = 0$ if $N(u, v) = 0$. We defined the topic-aware betweenness as

$$tb(w) = \frac{b(w)}{d(w, i)}. \tag{3}$$

Topic-aware Out-degree scoring method. Previous work has shown that simply removing nodes in order of decreasing out-degrees works well for preventing the spread of contamination in most real networks [4]. Here we focus on the contaminated set S and the corresponding $i \in \mathcal{I}$. We define the out-degree score $o(w)$ of node $w \in N(S)$ as the number of non-contaminative nodes around w. We defined the topic-aware out-degree as

$$to(w) = \frac{o(w)}{d(w, i)}. \tag{4}$$

Definition (3) and (4) are reasonable, since we can find that the smaller $d(w, i)$ is, the more susceptible the node w is; and the bigger $b(w)$ or $o(w)$ is, the more pivotal the node w is. Hence blocking the nodes with the highest topic-aware betweenness and outdegree score should be effective for preventing the spread of contamination in the network.

4. EXPERIMENT RESULTS

We experimentally evaluate the performance of our proposed approaches on two networks. One is crawled from Sina microblog containing 2,000 nodes, 14,426 edges and the propagation log. The other is Facebook data acquired from Stanford Network Analysis Project containing 4,039 nodes and 88,234 edges, where the topic probability for each user is created by the HDP-LDA model. We use the Gibbs sampling method to estimate the hyper parameters γ, α_0 and H in HDP-LDA. We employ the Monte-Carlo simulation of TIC model to estimate the influence spread.

From the results in Fig. 1, we can observe that the ultimate influence spreads by Topic-aware heuristics are significantly reduced compared to that by Out-degree and Betweenness centralities, especially in the early stage. For the infected set S with $|S| = 50$ on Sina microblog, we can observe that the proposed method can reduce the negative spread from 320 to 180 by blocking 60 nodes. Here the blocked 60 nodes only accounts to 15% of the nodes that are connected to infected nodes. Although the performance is improved greatly, the time cost of topic-aware heuristics are still in the same magnitude with centrality measures.

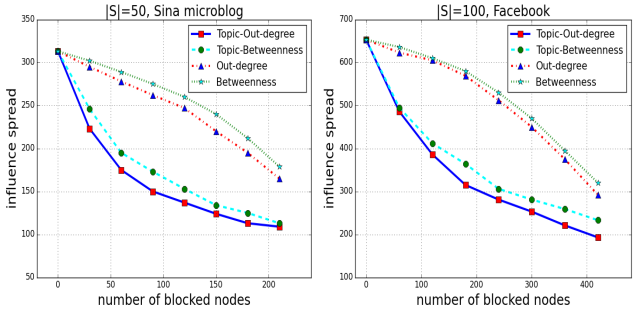

Figure 1: Experiment result on two data sets

Acknowledgement. This work was supported by National Grand Fundamental Research 973 Program of China under Grant No.2013CB329605, Chinese Universities Scientific Fund under Grant No.BUPT2014RC0701, National Science and Technology Support Program of China under Grant No.2013BAH43F00-01, and Strategic Leading Science and Technology Projects of CAS (No.XDA06030200).

5. REFERENCES

[1] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. In *Proc. ICDM 2012*, pages 81–90. IEEE Computer Society.

[2] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *Proc. WWW 2011*, pages 665–674. ACM.

[3] M. Kimura, K. Saito, and H. Motoda. Minimizing the spread of contamination by blocking links in a network. In *AAAI*, volume 8, pages 1175–1180, 2008.

[4] S. Wang, X. Zhao, Y. Chen, Z. Li, K. Zhang, and J. Xia. Negative influence minimizing by blocking nodes in social networks. In *AAAI (Late-Breaking Developments)*, 2013.