Influence Propagation and Maximization for Heterogeneous Social Networks

Cheng-Te Li[1], Shou-De Lin[1], Man-Kwan Shan[2]
[1] Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan
[2] Department of Computer Science, National Chengchi University, Taipei, Taiwan
{d98944005, sdlin}@csie.ntu.edu.tw, mkshan@cs.nccu.edu.tw

ABSTRACT

Influence propagation and maximization is a well-studied problem in social network mining. However, most of the previous works focus only on homogeneous social networks where nodes and links are of single type. This work aims at defining information propagation for heterogeneous social networks (containing multiple types of nodes and links). We propose to consider the individual behaviors of persons to model the influence propagation. Person nodes possess different influence probabilities to activate their friends according to their interaction behaviors. The proposed model consists of two stages. First, based on the heterogeneous social network, we create a human-based influence graph where nodes are of human-type and links carry weights that represent how special the target node is to the source node. Second, we propose two entropy-based heuristics to identify the disseminators in the influence graph to maximize the influence spread. Experimental results show promising results for the proposed method.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining*.

General Terms

Algorithms, Performance, Design.

Keywords

Influence propagation, influence maximization, social networks.

1. INTRODUCTION

Social network plays a key role for the spread of information and influence for viral marketing and online advertising. For example, advertisers may want to promote certain product via a social network. They have limited number of samples and aim to find a subset of initial users to test them, such that these users have higher potential to influence their friends if they like the product. Then through the word-of-mouth effect, a larger population in the network would adopt such product. This problem, termed *influence maximization* [3], is to find a set of influential individuals (called *seeds* or *disseminators*) such that they can eventually activate the largest number of people in a social network. So far some greedy [3][6] and heuristic [1][2] methods are proposed to effectively and efficiently solve this problem. Besides, there are some variations to tackle different real-world requirements. Leskovec et al. [5] 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. [4] proposes 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.

Existing proposals and solutions on influence maximization focus mainly on networks with only a single type of nodes and links. In this paper, we propose to model the influence propagation in *heterogeneous* networks which contains multiple typed labels on nodes and links. An example is shown in Figure 1. We believe through exploiting different types of objects and interactions/relationships, it is possible to design a more realistic and useful influence spreading mechanism.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

In this work, we consider the relational behaviors between nodes of human type to model the influence propagation in a heterogeneous social network. The basic idea is to transform the heterogeneous network to a probabilistic *influence graph*, in which nodes contains only human types and the weights on edges are probabilities derived by computing how regular the target node is to the source node using high-order relations (i.e., two persons are connected via one or more relations) between them. In addition, we consider two kinds of human interactions to model the influence propagation, the regular and rare behaviors. The former tends to spread the influence to those persons whose interaction behaviors are frequent with respect to the individual, while the latter is designed to activate the special friends (their interaction behaviors are rare w.r.t. the individual). We capture the interaction behaviors based on the high-order relations between persons, and thus determine the influence probability between persons by computing the relative frequency of the sequence of relation types, according to which kind of behavior is chosen. Using the modeled influence graph, we present two entropy-based heuristics to find the disseminators by influence maximization.

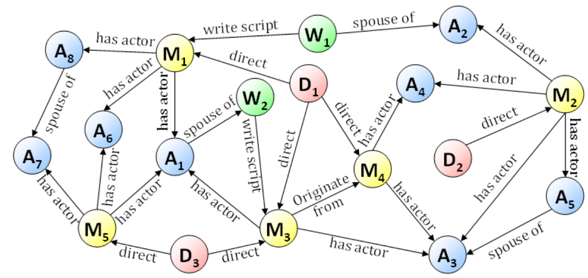

Figure 1. An example heterogeneous movie network, containing four types of nodes (Actor, Director, Writer, Movie) and five types of relations (WriteScript, HasActor, SpouseOf, Direct, OriginateFrom). Note that the node types A, D, and W are types of persons.

Preliminary. First, a *heterogeneous social network* $H=(V_H,E_H,L)$ is a directed labeled graph, where V_H is a set of nodes, L is a set of relation labels, and $E_H \subseteq V_H \times L \times V_H$ is a set of edges. Given a triple <source, label, target> representing an edge, the function $type(E) \rightarrow \{\{l_1,\ldots,l_j\}, l_i \in L, j \geq 1\}$ maps each edge e_H onto its set of typed labels. Note that the inverse edge set E^{-1} is the set of all edges (v,l_i^{-1},u) such that $(u,l_i,v) \in E$. Second, a *influence graph* $G=(V,E,W)$ is a bi-directed graph, where V is a set of nodes, E is a set of edges, and W is a weight function mapping each directed edge e to an *influence probability* $p(u \rightarrow v) \in [0,1]$. Third, we provide two kinds of behavior options, i.e., *'REGU'* and *'RARE'*, corresponding to the normal and special behaviors respectively. Fourth, we use the *Independent Cascade* (IC) model [3] for the information propagation. Let S be the set of seed nodes. In IC model, in time step t each active node u has a single chance to activate each of its inactive neighbors v with a certain probability $p(u \rightarrow v)$. If u succeeds, v will become active at step $(t+1)$. Otherwise, u will not activate v again. The *influence spread* of S is the expected number of activated nodes given seed set S.

Problem Statement. (1) Given a heterogeneous social network H, a range of local neighborhood r, and a behavior option $OP \in \{$*'REGU'*, *'RARE'*}, transform H into an influence graph $G=(V,E,W)$. (2) Given an influence graph G, a budget (integer) k, find a set of k nodes $S \subseteq V$ as the disseminators such that the influence spread is maximized using the independent cascade model.

2. THE PROPOSED METHOD

Our method consists of two stages: (1) behavior-based influence modeling, and (2) the entropy-based node selection.

Behavior-based Influence Modeling. Our model for influence propagation in a heterogeneous social network can be divided into two parts: (a) generating the influence probability from one person to another according to their relational interactions, and (b) constructing an influence graph containing nodes with human types (e.g. actor, director, and writer in a movie network). We exploit the relational structure in the neighborhood of each human node to model his/her personal behaviors. We first extract the r-step neighbor subgraph H'_v for each human node v. Constraining the size of the neighborhood is reasonable since it is usually assumed farer away nodes do not have as significant influence as the closer ones. Then, to capture the personal behavior of v, for each human node u in H'_v, we find the shortest path $path(v, u)$ and extract the *relation sequence* rs=<$l_1,\dots,l_d$>, where l_i (i=1,...,d) is the relation labels along $path(v, u)$ and d=|$path(v, u)$|. Then for each human node v, we can collect a set of relation sequences RS_v to describe the personal behaviors of v. For example, in Figure 1, the RS_{A1} is {<spouse of>, <HasActor^{-1}, HasActor>, <HasActor^{-1}, WriteScript^{-1}>, <HasActor^{-1}, Direct^{-1}>}. We add all the human nodes into the influence graph G. A directed edge ($v \rightarrow u$) connected v to u is established if and only if there the shortest path between them is smaller than r.

The second step of the behavior-based modeling is to compute the influence probability $p(v \rightarrow u)$ as the edge weight in G according to the specified behavior option OP and the extracted set of relation sequence RS. Since each outgoing edge ($v \rightarrow u$) of v in G represents a relation sequence $rs \in RS_v$ in H'_v, we can compute the number of times each rs occurs in H'_v. This number is denoted by $N_v(rs)$. Then based on the behavior option OP, we compute the contribution value for each $rs \in RS_v$ to capture how significant v performs the behavior rs. And this contribution value is used as the influence probability $p(v \rightarrow u)$. The influence probability is defined as

$$p(v \rightarrow u) = N_v(rs) / \Sigma_{x \in RSv} N_v(x), \quad \text{if } OP = \text{'REGU'},$$
$$p(v \rightarrow u) = 1 - (N_v(rs) / \Sigma_{x \in RSv} N_v(x)), \quad \text{if } OP = \text{'RARE'}.$$

In other words, if we specify the *'REGU'* as the option, one will have higher probability to activate those neighbors he/she frequently interacts with. We also can say a person tends to propagate information to its regular friends. If the option is *'RARE'*, it indicates people tend to have significant impact on those have special meanings (i.e. friends that it interacts differently) for them. For example, if OP=*'REGU'*, A_1 can affect other actors easily. But if OP=*'RARE'*, A_1 has higher potential to activate its spouse. The choosing of behavior options provides a flexible of influence propagation and can either depend on the nature of data or user's preference. Note that we normalize $p(v \rightarrow u)$ such that the sum of p for all outgoing edges of a node v equals to be 1.

Entropy-based Node Selection. Based on the influence graph G, we propose two entropy-based heuristic methods to select effective seed nodes for influence maximization in a heterogeneous network. The first utilizes the entropy value of the probabilistic weighs on outgoing edges of a node. Those with higher entropy values are selected as seeds. And in our behavior-based influence modeling, it is common for nodes to have more neighbors. The second heuristic is called *entropy discount* whose idea is based on *degree discount* [2]. The idea is that if an active node v successfully activates its neighbor u through the outgoing edge ($v \rightarrow u$), we should not take count the outgoing edge ($u \rightarrow v$) towards u's entropy. We should discount and recalculate u's entropy. The entropy discount is performed as each time activating an inactive node during the process of influence propagation.

3. EXPERIMENTAL RESULTS

We conduct the experiment on a real-world heterogeneous network. The goal is to demonstrate the usefulness of the proposed behavior-based influence modeling and two entropy-based heuristics for influence maximization in a heterogeneous social network. The UCI KDD movie data (http://kdd.ics.uci.edu/databases/movies/) is used to compile the heterogeneous network containing 39,212 nodes and 84,145 edges. There are 20 node types (11 types are about human) and 44 relation types. The modeled influence graph contains 25,943 nodes and 132,983 edges. To be fair, we compare our two entropy-based heuristics to two degree-based heuristics in [2][3]. To obtain the influence spread, for each selected seed set, we run the IC model on the influence graph 20,000 times and compute the average of the influence spread as the final score. Besides, we compare their final score by varying the size of selected seed set from k=1 to 51 and r is set as 2. The result is shown in Figure 2. We can find our entropy discount (EntropyDis) outperforms others, especially when k is large.

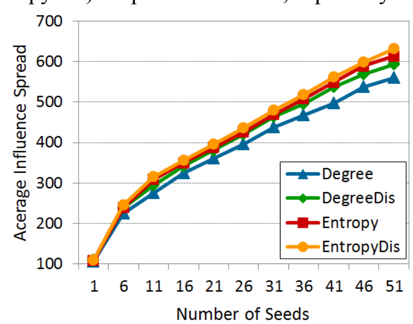

Figure 2. Experimental results by varying the number of seeds.

4. CONCLUSION

We propose to consider the personal behaviors of each individual to model the influence propagation in a heterogeneous network. Based on the given option of normal or special behavior, we compute the influence probability as weights between persons and derive the modeled influence graph. Besides, we present two entropy-based heuristics to select seed nodes as the disseminators in the influence graph. Experimental results on real-world movie data show the effectiveness of the proposed method.

5. ACKNOWLEDGEMENT

This work was supported by National Science Council, National Taiwan University, and Intel Corporation under Grants: NSC 100-2911-I-002-001, and 10R70501

6. REFERENCES

[1] W. Chen and C. Wang. Scalable Influence Maximization for Prevalent Viral. Marketing in Large-Scale Social Networks. In *KDD* 2010.

[2] W. Chen, Y. Wang, and S. Yang. Efficient Influence Maximization in Social Networks. In *KDD* 2009.

[3] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence through a Social Network. In *KDD* 2003.

[4] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding Effectors in Social Networks. In *KDD* 2010.

[5] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective Outbreak Detection in Networks. In *KDD* 2007.

[6] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based Greedy Algorithm for Mining Top-k Influential Nodes in Mobile Social Networks. In *KDD* 2010.