Exploiting Cluster-based Meta Paths for Link Prediction in Signed Networks

Jiangfeng Zeng, Ke Zhou, Xiao Ma✉, Fuhao Zou, Hua Wang
Huazhong University of Science and Technology
Wuhan China, 430074
{jfzeng, k.zhou, cindyma, fuhao_zou, hwang}@hust.edu.cn

ABSTRACT

Many online social networks can be described by signed networks, where positive links signify friendships, trust and like; while negative links indicate enmity, distrust and dislike. Predicting the sign of the links in these networks has attracted a great deal of attentions in the areas of friendship recommendation and trust relationship prediction. Existing methods for sign prediction tend to rely on path-based features which are somehow limited to the sparsity problem of the network. In order to solve this issue, in this paper, we introduce a novel sign prediction model by exploiting cluster-based meta paths, which can take advantage of both local and global information of the input networks. First, cluster-based meta paths based features are constructed by incorporating the newly generated clusters through hierarchically clustering the input networks. Then, the logistic regression classifier is employed to train the model and predict the hidden signs of the links. Extensive experiments on Epinions and Slashdot datasets demonstrate the efficiency of our proposed method in terms of Accuracy and Coverage.

Categories and Subject Descriptors

H.3.3 [**Information Search Retrieval**]: Information Filtering; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Signed Networks; Link prediction; Cluster-based Meta Path

1. INTRODUCTION

The increasing popularity of social networks allows online users to participate in online activities conveniently. Users can form positive links with others through friendships in social network websites, e.g., Facebook, Twitter, LinkedIn, etc, as well as establish negative relationships, e.g., distrust, foes, in websites like Epinions, Slashdot, etc.

In order to help users better connect to others, a set of link prediction methods have been explored to detect the latent relationship in the networks which may link two users together in the future[6, 7, 9]. Existing link prediction algorithms can be roughly divided into two groups, which correspond to supervised and unsupervised methods. Supervised methods consider the link prediction problem as a classification problem by using the existence of links as labels, while unsupervised methods make use of the topological properties of the snapshot of the network[7, 11].

Typically, link prediction methods only consider the presence/absence of a connection between two nodes, and ignore the possible sign of the connection. However, in signed networks, not only the link prediction problem is important, but also correctly predicting the sign of the links matters. Existing sign prediction algorithms mainly rely on path-based features, and the effectiveness of these models depend on the number of paths between the target nodes[1, 5, 8]. That is to say, the accuracy of prediction is higher for edges with larger values of embeddedness (embeddedness is the number of paths with length two between the target nodes). However, the sparsity problem is serious in real networks. As a result, it will be quite difficult to find such kind of paths connecting arbitrary nodes in such sparse networks.

Our proposed method for the problem is motivated by the meta paths based relationship prediction model proposed in heterogeneous information networks[10]. Different meta paths capture different semantics between any two vertices in the networks, which consider not only the regular paths between nodes, but also the structural information of the networks. It has been verified by Sun et al.[10] that the prediction accuracy of meta paths based models are usually higher than other traditional path based models, e.g., Personalized-PageRank[2], SimRank[3]. By extending their intuition into the homogeneous networks, we assume that two nodes are more likely to be similar in the networks if their neighbors belong to the same clusters. Based on this intuition, we introduce the cluster-based meta paths, which can efficiently alleviate the sparsity problem appeared in pure path based sign prediction models by employing not only the local information of nodes, but also the global structural information of the networks.

The contributions of our work can be summarized as follows. **First**, we propose a novel cluster-based meta path which incorporates meaningful network structural information into path-based features for link prediction in signed networks. **Second**, we show the effectiveness of cluster-based meta path features by thoroughly comparing the pre-

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983870

diction results generated by pure path-based features and cluster-based meta path features of different cluster layers.

The rest of the paper is organized as follows: the notations are defined in Section 2. Section 3 introduces the proposed sign prediction model. The experiment results and analysis are presented in Section 4. Finally, Section 5 concludes this study with future work.

2. PRELIMINARY

A signed network can be denoted as a graph $G = (V, E, \Sigma)$ where $V = \{1, 2, ..., |V|\}$ is a finite set of nodes of the graph. For $\forall u_i, u_j \in V$, E consists a set of directed or undirected edges $e_{i,j}$ between two nodes. The third component of the graph is a mapping $\Sigma : s_{e_{i,j}} \rightarrow \{+1, -1\}$ giving a sign to each edge, where $s_{e_{i,j}} = 1$ represents that positive relationship exists between node u_i and node u_j, while $s_{e_{i,j}} = -1$ means that these two nodes share a negative relationship in the graph.

3. THE APPROACH

3.1 Problem definition

Formally, given a graph $G = (V, E, \Sigma)$, and a test edge $e_{i,j} \in E$ whose sign is hidden, the problem is to infer the disappeared sign of the edge $e_{i,j}$ based on the information extracted from the rest of the graph.

3.2 Meta Paths

In a heterogeneous information network schema, two entities can be connected via different paths, which usually carry different semantic meanings. In this paper, we use the meta path definition from [10] in a network schema as follows:

Definition 1. **Meta Path(MP)**

A meta path $P = A_0 \xrightarrow{R_1} A_1 \xrightarrow{R_2} ... \xrightarrow{R_l} A_l$ is a path defined on the graph of network schema $T_G = (\mathcal{A}, \mathcal{R})$, and defines a new composite relation $R_1 R_2 ... R_l$ between type A_0 and A_l, where $A_i \in \mathcal{A}$ and $R_i \in \mathcal{R}$ for $i = 0, ..., l$. $A_0 = dom(R_1)$, $A_l = range(R_l)$, and $A_i = range(R_i) = dom(R_{i+1})$ for $i = 1, ..., l-1$.

Different meta paths usually contain different semantics in a heterogeneous graph. The following examples are two representative meta paths in the DBLP network schema:

$P_1 : paper \xrightarrow{PublishedIn} venue \xrightarrow{PublishedIn^{-1}} paper$

$P_2 : \ paper \xrightarrow{Contains} term \xrightarrow{Contains^{-1}} paper$

Meta path P_1 defines the relationship of publishing in the same venue for two papers. Intuitively, if two papers published in the same venues, these two papers seem to belong to the same research area. P_2 captures the relationship of sharing similar research topics between two papers. It can be seen that meta paths can capture different relationships hidden in the network, which are extremely helpful to capture the structural information embedded in a heterogeneous graph.

Now let's think about the role of entity 'venue' in the above meta paths. Many papers are published in the same venues, therefore, 'venue' can be considered as hubs which have the attribute of attracting large volume of other types of entities in the networks. Inspired by this phenomenon, we begin to think that whether it is possible to introduce a type of entities which can be considered as hubs in homogeneous networks. Intuitively, a cluster containing multiple nodes is exactly qualified for this requirement. Therefore, in the setting of homogeneous networks, we first introduce the definition of node-based meta paths as follows, and the cluster-based meta paths will be discussed in the next section.

Definition 2. **Node-based Meta Path(NBMP)**

A node-based meta path is defined as $U \xrightarrow{R} U \xrightarrow{R^{-1}} U$, where $U \in V$, R represents composite relations between any pair of nodes in the networks, and the link type R^{-1} is the inverse of relation R. Simplest instances of node-based meta paths are the paths with length two, which can be denoted by $p : u_i - u_k - u_j$, where $u_i, u_k, u_j \in V$, and $e_{i,k}, \ e_{k,j} \in E$.

3.3 Cluster-based Meta Paths based Sign Prediction Model

The proposed model can be executed in three steps. **First**, in order to investigate how the network structure influence the quality of meta-paths, we hierarchically clustering the input network and adding a set of new entities into the input network based on the newly obtained clusters of different layers. **Second**, defining features based on meta-paths which include the entities of clusters. **Third**, predicting the signs of links using the logistic regression classifier.

Note that, in the first step, plenty of algorithms can be used for the hierarchical clustering task. In this paper, we employ the InfoMap clustering algorithm proposed in [4], where the input is the signed networks with both positive and negative links, and the output is the hierarchical clusters. Based on the newly obtained clusters, we introduce the definition of cluster-based meta paths as follows:

Definition 3. **Cluster-based Meta Path(CBMP)**

A cluster-based meta path is defined as $U \xrightarrow{R} U \xrightarrow{B} C_x \xrightarrow{B^{-1}} U \xrightarrow{R^{-1}} U$, where $U \in V$, R represents composite relations between two arbitrary nodes. C_x signifies the clusters in the x^{th} layer. B denotes a node belongs to cluster C_x, and B^{-1} is the inverse of relation B. $C_{x,y} (C_{x,y} \in C_x)$ means the y^{th} cluster in layer x.

Figure 1 is a simple example describing how cluster-based entities can be added to the input network. As we can see that, 3 first-layer clusters $C_{1,1}$, $C_{1,2}$,$C_{1,3}$, as well as 2 second-layer clusters $C_{2,1}$, $C_{2,2}$ are obtained through clustering and added to the network as some super nodes. For example, nodes u_1, u_2, u_3, u_i belong to the first-layer cluster $C_{1,1}$; and nodes u_9, u_{11} belong to the first-layer cluster $C_{1,3}$ and second-layer cluster $C_{2,1}$. The solid links represent the relationships with known labels between nodes in the input network, and the dotted links denote edges with hidden signs (For simplicity, the directions and signs of links are omitted.).

Suppose the nodes u_i and u_j are two target nodes, the neighbors of node u_i are u_1 and u_9, and the neighbors of node u_j are u_2, u_5, u_7, u_8 and u_{11}. According to Definition 2, only 2 instances of node-based meta paths exists in this graph, that is, $p_1 : \ u_i - u_1 - u_2 - u_j$; $p_2 : \ u_i - u_9 - u_{11} - u_j$. However, according to Definition 3, there are 4 instances of the cluster-based meta paths existing in the first layer, and 1 instance in the second layer: $p_{11} : u_i - u_1 - C_{1,1} - u_2 - u_j$; $p_{12} : u_i - u_9 - C_{1,3} - u_7 - u_j$; $p_{13} : u_i - u_9 - C_{1,3} - u_8 - u_j$; $p_{14} : u_i - u_9 - C_{1,3} - u_{11} - u_j$; $p_{21} : u_i - u_9 - C_{2,1} - u_{11} - u_j$.

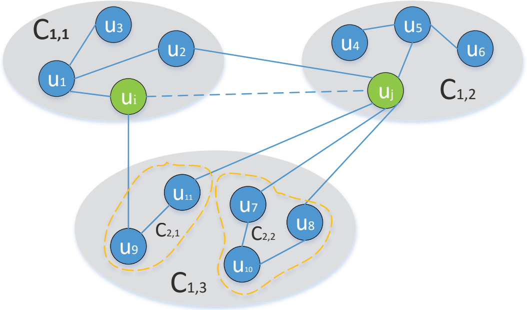

Figure 1: A toy example describing the relations between nodes and clusters.

In fact, paths p_{11}, p_{14} and p_{21} are equivalent to node-based meta paths p_1 and p_2. However, there are two more cluster-based meta paths, e.g., p_{12}, p_{13}, which include the structural information of the network by exploiting clustering. In other words, our proposed cluster-based meta paths can guarantee having sufficient number of paths between nodes, which will effectively alleviate the sparsity problem.

Feature computation. The values of CBMP features can be computed by the number of path instances following meta paths of different cluster layers between the target nodes u_i and u_j. For example, the feature vector for edge $e_{i,j}$ in Figure 1 can be denoted as $f_{i,j} = \{f_1, f_2\}$, where $f_1 : 1 \times 1 + 0 + 1 \times 3 = 4$, $f_2 : 1 \times 1 + 0 = 1$.

Note that, it is more efficient to employ matrix multiplication to find the path instances than directly searching paths in the whole dataset. Two types of matrices are essential for the multiplication in our setting: one is the *Node-Node Adjacency Matrix*; the other is the *Node-Cluster Belong-ship Matrix* representing the relationships between each node and each cluster of different cluster layers. Due to limited space, we will not discuss it further. Please refer to [10] for more detail.

Learning methodology. Following the method in [5, 8], we apply the logistic regression classifier to do edge sign prediction based on both the node-based meta paths features and cluster-based meta paths features. Logistic regression learns a model as follows:

$$p^+(s_{e_{i,j}}) = \frac{1}{1 + e^{-(a_0 + \sum\limits_{m=1}^{k} a_m f_{i,j}^m)}} \tag{1}$$

The coefficients $\{a_0, ..., a_k\}$ are learned from the training data using maximum likelihood estimation, and k is the number of features. If $p^+(s_{e_{i,j}}) \geq 0.5$, there exists a positive relationship between nodes u_i and u_j, else negative results are obtained.

4. EXPERIMENTS AND ANALYSIS

4.1 Datasets Description

Two real-world online social networks are used in the experiments, namely Epinions and Slashdot downloaded from *http://snap.stanford.edu/data/*. Both the networks have explicit sign labels on the links. By preprocessing the datasets, we filter out the nodes which only exist in positive links or negative links. The statistics of datasets we used in our experiment are shown in Table 1. From this table we can conclude that both the datasets are quite sparse and imbalanced, i.e., more positive links exist than negative links.

Table 1: Statistics of the datasets

	Epinions	*Slashdot*
# of nodes	114,476	75,144
# of links	807,894	537,665
Fraction of + links	88.83%	79.06 %
Fraction of - links	11.17%	20.96%
Sparsity	99.38%	99.04%

4.2 Methodology and Metrics

We evaluate the proposed method using a *leave-one-out* methodology: each edge in the testing network is successively removed and the methods try to predict the sign of that edge using the rest of the network. In order to avoid over-fitting, we adopt 10-*fold cross-validation*. We randomly created 10 disjoint test folds each consisting of 10% of the whole number of edges in the networks. For each test fold, the remaining 90% of the links serve as the training set. Then the accuracy and coverage are generated by averaging them over the 10 folds.

Since the datasets are quite imbalanced, we follow the methodology of [5] and create a balanced dataset with equal numbers of positive and negative edges, so that random guessing yields a 50% accuracy. In order to assess the performance of the methods, we computed the accuracy of the prediction results as $Accuracy = \frac{CE}{|E|}$, where CE and $|E|$ represent the number of edges for which the sign is correctly predicted and the total number of edges in the network, respectively. Another evaluation metrics we are interested in is the prediction coverage: $Coverage = \frac{PE}{|E|}$, where PE is number of predictable edges. The coverage metric measures the degree to which the missing data can be predicted and covered relative to the whole dataset.

4.3 Results Analysis

In this study, 5 groups of meta paths based features have been used when implementing the proposed model:

$p_1 : U \overset{R}{\to} U \overset{-R}{\to} U$, $p_2 : U \overset{R}{\to} U \overset{B}{\to} C \overset{B^{-1}}{\to} U \overset{R^{-1}}{\to} U$,

$p_{3(4,5)} : U \overset{R}{\to} U \overset{B}{\to} C_{1(2,3)} \overset{B^{-1}}{\to} U \overset{R^{-1}}{\to} U$

In order to decrease the complexity, we treat the composite relations R as the adjacency relations, therefore, p_1 is degraded to the simplest path-based feature with length two; p_2 is a special type of cluster-based meta path where the whole dataset is considered as a super cluster; p_3, p_4 and p_5 are cluster-based meta paths of the first cluster layer, second cluster layer and third cluster layer, respectively. The reason for selecting the first three cluster layers can be found in Section 4.4. Note that each R can have two directions (inlink, outlink) and two signs (positive, negative), that is to say, each type of meta paths feature is composed of 16 sub-features. Thus we have $16 \times 5 = 80$ features in all.

In order to demonstrate the efficiency of our proposed cluster-based meta paths features, we compare the prediction accuracy and coverage metrics of the sign prediction models based on different features. SBP represents social balance based features which are proposed in [5], and AllMP denotes a combination of node-based meta paths features and cluster-based meta paths features.

From Figure 2 we can observe that, pure path-based features, e.g., SBP and NBMP, perform a little bit better than

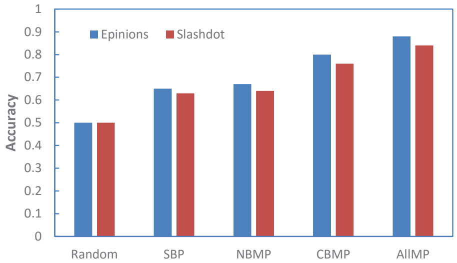

Figure 2: Accuracy of the predictor based on different meta paths features.

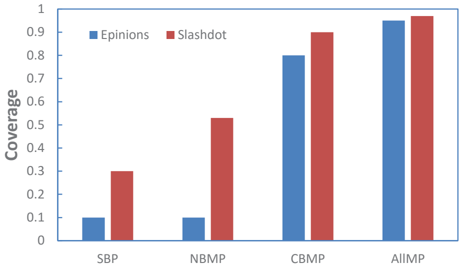

Figure 3: Coverage of the predictor based on different meta paths features.

random guessing, which is consistent with the results reported in [5]. In our consideration, this is mainly caused by the sparsity problem of networks. Not enough balanced triads or node-based paths can be found in such sparse datasets, which largely influence the accuracy of the prediction results. CBMP achieves better results than NBMP. The reason is that CBMP captures not only the local information, but also the global topological features by graph clustering as we have discussed in Section 3.3. In addition, consistent with our intuition, the best prediction accuracy is achieved by AllMP features, which verifies the fact that any one type of features cannot singly produce the best result.

Figure 3 shows the coverage. CBMP and AllMP yield better coverage rate than SBP and NBMP, which demonstrate that our proposed cluster-based meta paths features can effectively alleviate the sparsity problem in real networks.

4.4 Effect of hierarchical clustering layers

Figure 4 describes how the prediction accuracy is influenced by cluster-based meta paths of different cluster layers. When the clustering layer is 0, the whole dataset is considered as one big cluster, substantial meta paths can be found, which makes the final result inaccurate. When the clustering becomes fine-grained, the results get better. This can be explained as informative network structures are generated with the fine-grained clustering. As we can see that, the results are getting stable when the cluster layer reaches 3. This is because the sparsity problem is serious in much too fine-grained clusters, therefore, few cluster-based meta paths exist in these cluster layers.

5. CONCLUSION

In this paper, we have introduced the cluster-based meta paths based model which can effectively alleviate the sparsity problem in sign prediction tasks. By hierarchically clustering the input networks, newly generated clusters are added to the networks and incorporated into cluster-based meta paths. The logistic regression classifier is employed to predict the hidden signs of links based on the proposed meta paths features of different cluster layers. Extensive comparative experiments demonstrate the efficiency of our model.

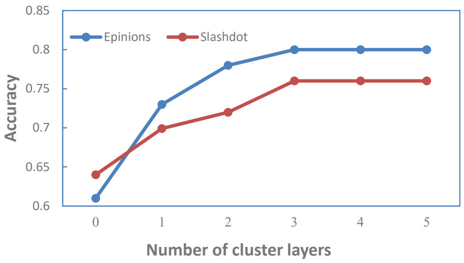

Figure 4: Accuracy of the predictor based on CBMP of different cluster layers.

One of the main issues in defining meta paths is how to set the length of meta-paths, in this paper, we set the composite relations R as the adjacency relations for simplicity. In the future, we will find out how to set a proper length of cluster-based meta paths. Besides, it is interesting to see how the accuracy will be like if different kinds of clustering methods are applied.

6. ACKNOWLEDGMENTS

This research is funded by the National Natural Science Foundation of China under grants No.61232004 and No.61502189, and the Fundamental Research Funds for the Central Universities, HUST:2016YXMS088.

7. REFERENCES

[1] K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon. Exploiting longer cycles for link prediction in signed networks. In *CIKM 2011*.

[2] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, 2(3):333–358, 2005.

[3] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD 2002*.

[4] A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[5] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW 2010*.

[6] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[7] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A*, 390(6):1150–1170, 2011.

[8] X. Ma, H. Lu, and Z. Gan. Implicit trust and distrust prediction for recommender systems. In *WISE 2015*, pages 185–199. Springer, 2015.

[9] T. Murata and S. Moriyasu. Link prediction based on structural properties of online social networks. *New Generation Computing*, 26(3):245–257, 2008.

[10] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB*, 2011.

[11] J. Tang, Y. Chang, C. Aggarwal, and H. Liu. A survey of signed network mining in social media. *arXiv preprint arXiv:1511.07569*, 2015.