Link Prediction in Social Networks Based on Hypergraph

Dong Li, Zhiming Xu, Sheng Li, Xin Sun
Harbin Institute of Technology
Harbin, 150001, P.R. China
{hitlidong, xuzm, lisheng, sunxin}@hit.edu.cn

ABSTRACT

In recent years, online social networks have undergone a significant growth and attracted much attention. In these online social networks, link prediction is a critical task that not only offers insights into the factors behind creation of individual social relationship but also plays an essential role in the whole network growth. In this paper, we propose a novel link prediction method based on hypergraph. In contrast with conventional methods that using ordinary graph, we model the social network as a hypergraph, which can fully capture all types of objects and either the pair wise or high-order relations among these objects in the network. Then the link prediction task is formulated as a ranking problem on this hypergraph. Experimental results on Sina-Weibo dataset have demonstrated the effectiveness of our methods.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

link prediction; hypergraph; ranking

1. INTRODUCTION

With the rapid development of networking sites (e.g., Facebook, Twitter and LinkedIn), online social networks have drawn substantial attention. Link prediction is the problem of predicting the existence of a link between two entities, based on attributes of the objects and other observed links. It is a critical task for social networks analysis, and has many applications, such as, user recommendation, network growth modeling and so on.

Nowell and Kleinberg [2] propose an array of methods for link prediction using network topology. They model a social network as a homogeneous graph, in which, each node represents a user and each link denotes social relationship between users. However, a social network usually does not only contain user objects and relationship between them, Fig. 1 presents an example of various types of objects and relations among these objects in micro-blogs. As an improvement, Yin et al. [4] model social networks as heterogeneous graphs and apply a random walk algorithm on them to calculate link proximity. But, their heterogeneous graphs only consider user objects and user attributes objects, they still can not capture completely all types of objects and relations in social networks. Moreover, the heterogeneous graphs they used are ordinary graphs, which can not make full use of the high-order relations in social networks such as R5 in Fig1 represents a ternary relation (a user releases a tweet containing a hashtag). For an ordinary graph, naively squeezing the complex relationships into pair wise ones will inevitably lead to loss of information [5].

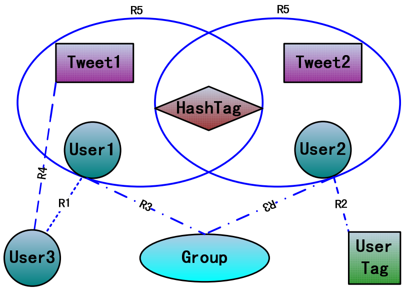

Figure 1. An example of various types of objects and relations among these objects in micro-blogs.

In this paper, we propose a novel link prediction method based on hypergraph. Hypergraph model has been widely used such as solving the problems of community detection [3], classification [5], music recommendation [1] and so on. But to the best of our knowledge, none has considered using hypergraph model to solve link prediction problem. In this paper, we use a hypergraph to model all types of objects and relations of the social network. This method, particularly, can fully capture the high-order relations among. Then we use a hypergragh ranking method for link proximities estimation and so as to make accurate link prediction. Experiments on Sina-Weibo dataset show that our methods outperform the state-of-the-art.

2. METHODS

A hypergraph is a generalization of an ordinary graph where edges, called hyperedges, can connect any number of nodes. Formally, let $G(V,E,w)$ denote a hypergraph, where V denotes a finite set of nodes *v*, *E* denotes the set of hyperedges *e*, *w* is a weight function defined as $w: E \rightarrow \mathbb{R}$. Each hyperedge $e \in E$ is a subset of *V* and is assigned a positive weight *w(e)*. The degree of a hyperedge e is defined as $\delta(e) = |e|$. For a node $v \in V$, the degree of v is defined as $d(v) = \sum_{e \in E|v \in e} w(e)$. A hypergraph G can be represented by a $|V| \times |E|$ matrix H with entries $h(v,e) = 1$ if $v \in e$ and 0 otherwise, called the incidence matrix of G. Then we have: $d(v) = \sum_{e \in E} w(e)h(v,e)$ and $\delta(e) = \sum_{v \in V} h(v,e)$. Let D_v and D_e denote the diagonal matrices containing the degrees of all nodes and hyperedges respectively, and W the diagonal matrix $|E| \times |E|$ containing the weights of all hyperedges.

We use a hypergraph to model different types of objects and either the pair wise or high-order relations among them in social networks. We suppose this hypergraph contains *m* kinds of nodes and *n* kinds of hyperedges. Each kind of node corresponds to a type of object, and each kind of hyperedge corresponds to a type of relation. A hyperedge in this hypergraph can be a set of nodes of either the same type or different types.

We consider link prediction task as a ranking problem on the above hypergraph. Given a target user *u*, we estimate link

WWW 2013 Companion, May 13–17, 2013, Rio de Janiero, Brazil.
ACM 978-1-4503-2038-2/13/05.

proximities between u and other nodes in the hypergraph for link prediction. Let $y=[y_1,y_2,\cdots,y_{|v|}]^T$ denote the input vector where y_i is the initial link proximity between y_i and the target user u. We use $f=[f_1,f_2,\cdots,f_{|v|}]^T$ to represent the result vector of link proximities. Next, we will discuss how to perform ranking based on hypergraph to estimate the link proximities *f* for link prediction. For a hypergraph ranking problem, the cost function $\Omega(f)$ could be defined as:

$$\Omega(f)=\frac{1}{2}\sum_{i,j=1}^{|V|}\sum_{e\in E}\frac{1}{\delta(e)}\sum_{\{v_i,v_j\}\subseteq e} w(e)\left\|\frac{f_i}{\sqrt{d(v_i)}}-\frac{f_j}{\sqrt{d(v_j)}}\right\|^2 \tag{1}$$

The function $\Omega(f)$ sums the changes of the scoring vector *f* over the hyperedges on the hypergraph. On the one hand, a good result vector *f* should make $\Omega(f)$ as small as possible, i.e. if two users have many same neighbors, the link proximity between them should be high. On the other hand, the initial score assignment should be changed as little as possible. Then the optimal ranking result is achieved by solving the following optimization problem:

$$\arg\min_{f\in R^{|V|}}\{\Omega(f)+\mu\|f-y\|^2\} \tag{2}$$

where $\mu>0$ is the parameter specifying the tradeoff between the two competitive terms. The calculation process is omitted due to space limitation, and finally we can obtain a closed-form solution,

$$f^*=(1-\alpha)(I-\alpha\mathrm{M})^{-1}y \tag{3}$$

Where $\alpha=1/(1+\mu)$ and $M=D_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}$.

3. EXPERIMENTS

To evaluate the proposed methods, we collected a dataset from Sina Weibo which is Twitter of China and now has more than 500 million users. This dataset contains various types of objects and relations shown in Fig.1. Firstly, we selected 15 seed users that are related to the internet field. Then we collected users followed by seed users and the “following” relations among all these users. Secondly, we collected usertags marked to above users, groups (called micro-groups in Sina Weibo) these users joined, and tweets republished or forwarded by these users. The content of tweet may contain hashtags or url, here we consider a url as a special hashtag. Finally, we collected 3910 users, 235790 tweets, 6080 usertags, 4462 hashtags and 593 groups. And the description and count of relations collected are summarized in Table 1. Each relation corresponds to a hyperedge, we set the weights of all hyperedges to be 1. Due to the revision of the Sina Weibo, the group information can not be collected any more now.

Table 1. Relations in dataset of Sina Weibo.

Notations	**Relation Description**	**Count**
R1	Users follow other users	**23213**
R2	Users assign usertags to themselves	**14138**
R3	Users join groups	**9762**
R4	Users forward tweet	**129729**
R5	Users release tweets containing hashtags or not	**106061**

We split the collected Sina Weibo dataset into training dataset and testing dataset. The training dataset is used to construct the hypergraph while the testing dataset is for evaluation. For each user in testing dataset, we remove half links to friends, and the prediction task is then to use the pruned networks to find the missing links. We compare the proposed method with several state-of-the-art methods: SimProf is to predict links using similarity of user profile, each user is profiled as a word vector based on tweets associated with this user. CommonNeighbors, Katz and Adamic/Adar are methods based on graph structure in [2]. LINKREC [4] uses a random walk algorithm on an augmented social graph with both structure and attribute information. We use Precision and Recall as the measure metrics to evaluate the performance of all the compared methods.

Table 2. Comparison of different methods on Sina Weibo dataset.

Method	**P@1**	**P@5**	**P@10**	**P@15**	**Recall**
SimProf	0.89	0.76	0.65	0.43	30.32
Common neighbors	4.87	4.69	4.50	4.25	47.46
Katz	4.86	4.66	4.49	4.21	45.90
Adamic/Adar	4.91	4.72	4.53	4.29	48.35
LINKREC	5.73	5.46	5.23	4.97	56.71
Our Method	**6.68**	**6.32**	**6.04**	**5.76**	**62.95**

Table 2 shows the results of these methods on Sina Weibo dataset. Common Neighbors, katz and Adamic/Adar show similar performance while SimProf gets a comparatively worse result. This indicates that network structure mining is more important than user profile analyzing in the task of link prediction. LINKREC outperforms all above motioned methods. However, LINKREC only focus on the graph consists of users and users’ attributes (usertags in micro-blog), it neglects other objects (tweets, groups and so on) and relations between users and these objects. In contrast, our method, modeling a social network as a hypergraph, can completely capture all types of objects and relations (either pair wise or high-order relations) in the social network, thus can get a better performance than LINKREC. Finally, we note that although this paper uses micro-blog dataset in evaluation, the proposed method is not limited to the micro-blog, it is also applicable to any other social network.

4. CONCLUSION

In this paper, a novel link prediction method based on hypergraph has been presented. In contrast with traditional methods based on homogeneous or heterogeneous graph, we model a social network as a hyperpgraph, which can accurately capture various types of objects and either the pair wise and high-order relations among these objects without loss of any information. Then we formulate the link proximity estimation task in link prediction as a ranking problem on this hypergraph. Experimental results on Sina Weibo dataset have proved the effectiveness of our methods.

Acknowledgements. This work is supported by the Natural Science Foundation of China (No. 61173074) and the ZTE cooperation project (No. MH20120428).

5. REFERENCES

[1] Bu, J., Tan, S., Chen, C., et al. Music recommendation by unified hypergraph combining social media information and music content. *In Proc. MM*, pages 391-400, 2010.

[2] Liben-Nowell, D. and Kleinberg, J. The link prediction problem for social networks. In Proc. CIKM, pages 556-559, 2003.

[3] Lin, Y. R., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A. Metafac: community discovery via relational hypergraph factorization. *In Proc. KDD*, pages 527-536, 2009.

[4] Yin, Z., Gupta, M., Weninger, T., and Han, J. LINKREC: a unified framework for link recommendation with user attributes and graph structure. In *Proc. WWW*, pages 1211-1212, 2010.

[5] Zhou, D., Huang, J., and Scholkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *In Proc. NIPS*, pages 1601-1608, 2007.