LINKREC: A Unified Framework for Link Recommendation with User Attributes and Graph Structure*

Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801
{zyin3, gupta58, weninge1, hanj}@illinois.edu

ABSTRACT

With the phenomenal success of networking sites (*e.g.*, Facebook, Twitter and LinkedIn), social networks have drawn substantial attention. On online social networking sites, link recommendation is a critical task that not only helps improve user experience but also plays an essential role in network growth. In this paper we propose several link recommendation criteria, based on both *user attributes* and *graph structure*. To discover the candidates that satisfy these criteria, link relevance is estimated using a random walk algorithm on an augmented social graph with both attribute and structure information. The global and local influence of the attributes is leveraged in the framework as well. Besides link recommendation, our framework can also rank attributes in a social network. Experiments on DBLP and IMDB data sets demonstrate that our method outperforms state-of-the-art methods based on network structure and node attribute information for link recommendation.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

link recommendation, random walk

1. INTRODUCTION

Social networking sites, such as Facebook, Twitter, and LinkedIn, are increasingly popular. Facebook reports that it has 300 million active users, 50% of whom login every day on average. Worldwide, more than 8 billion minutes are spent on Facebook per day. They use the social network sites not only to maintain contacts with old friends, but also to find new friends with similar interests and for business networking. It is reported that an average user has 130 friends on Facebook. Since the linkage among people is the underlying key concept for online social network sites, it is not surprising that link recommendation is an essential link mining task [1].

In this paper, we propose a framework using both attribute and structural properties to recommend potential linkages in social networks. To compute accurate link recommendations in social networks, we propose a list of desired criteria. A random walk framework on an augmented social graph using both attribute and structural properties is further proposed, which satisfies all the criteria. We also discuss different methods for setting edge weights in the augmented social graph which considers both global and local characteristics of the attributes. The experiments on two real data sets show that our methods outperform the state-of-the-art methods.

* Research supported in part by NSF under grant IIS-09-05215 and the Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. PROBLEM FORMULATION

Given a *social graph* $G(V, E)$, where V is the set of nodes and E the set of edges, each node in V represents a person in the network, and each edge in E represents a link between two person nodes. Besides the links, each person has his/her own attributes.

The *link recommendation task* can be expressed as follows: Given node v in V, provide a ranked list of nodes in V other than the existing linked nodes of v as the potential links ranked by link relevance. The following presents some intuition-based desiderata for link relevance: (1) *homophily*: two persons who share more attributes are more likely to be linked than those who share fewer attributes; (2) *rarity*: the rare attributes are likely to be more important, whereas the common attributes are less important; (3) *social influence*: the attributes shared by a large percentage of friends of a particular person are important for predicting potential links for that person; (4) *common friendship*: the more friends two persons share, the more likely they are to be linked together; (5) *social closeness*: the potential friends are likely to be located close to each other in the social graph; and (6) *preferential attachment*: a person is more likely to link to a popular person rather than to a person with few friends.

A good link candidate should satisfy the above criteria in social networks. In other words, the link relevance should be estimated by considering these intuitive rules.

3. LINKREC FRAMEWORK

In order to recommend potential links, we augment the original social graph with attribute nodes. To discover link candidates based on the criteria in Section 2, we design a random walk based algorithm on the augmented graph and use it to compute the link relevance. For each given person node, we rank all the candidates by link relevance, and suggest the top few for link recommendation.

Given the original social graph $G(V, E)$, we construct a new augmented graph $G'(V', E')$. Specifically, for each node in graph G, we create a corresponding node in G', called *person node*. For each edge in E in graph G, we create a corresponding edge in G'. For each attribute a, we create an additional node in G', called *attribute node*. For every attribute of a person, we create a corresponding edge between the person node and the attribute node. The edge weights from attribute nodes to person nodes are defined in a uniform way. The edges starting from person nodes can point to either person nodes or attribute nodes; parameter λ controls the tradeoff of the weights between these two types. The larger the λ is, the more the algorithm uses attribute properties for link recommendation. For the edges from person nodes to attribute nodes, we have

several weighting options: (1) weight all the attributes equally for each person; (2) attach more weight to the more globally important attributes, where the global importance of attribute measures the percentage of existing links among all the possible person pairs with that attribute; (3) attach more weight to the more locally important attributes, *i.e.,* the more the number of friends that share the attribute, the more important the attribute is for the person; and (4) mix global and local importance together by linear interpolation or multiplication.

In order to calculate the link relevance based on these criteria, we propose a random walk based algorithm on the newly constructed graph to simulate the friendship hunting behavior. We calculate the link relevance with regard to a particular person by using random walk with restart [3]. Random walk process on the newly constructed graph satisfies the desiderata for link relevance in the following ways: (1) *homophily*: if two persons share more attributes, it is more likely to walk from one person to the other; (2) *rarity*: if one attribute is rare, there are fewer outlinks for the corresponding attribute node, because the weight of each outlink is larger, so the probability of a random walk originating from a person via this attribute node is larger; (3) *social influence*: if one attribute is shared by many of the existing linked persons of the given person, the random walk is more likely to pass through the existing linked person nodes to this attribute node; (4) *common friendship*: if two persons share many friends, it is more likely to walk from one person to the other; (5) *social closeness*: if two persons are close to each other in the graph, the random walk probability from one to the other is likely to be larger than if they are far away from each other; and (6) *preferential attachment*: if a person is very popular, there are many inlinks to the person node in the graph, and for a random person node in the graph, it is easier to access a node with more inlinks.

Besides link recommendation, we can rank attributes with respect to a specific person by using the proposed framework. Attribute ranking can have many potential applications. For example, advertisements can be targeted more accurately if we know a person's interests more precisely. In the augmented graph, all the nodes including the attribute nodes have the random walk probability, so we can rank attribute nodes based on the random walk probability. The attributes with high ranks in our framework are those that can be easily accessed by not just the given person but also the existing friends and the potential friends.

Besides link recommendation for a single person, we can also recommend links for a community by restarting the random walk from nodes belongs to one community instead of a single one. Similarly, we can also rank the attributes for a cluster of person nodes.

4. EXPERIMENTS

We conduct our experiments on two real data sets to demonstrate the effectiveness of our framework. For DBLP[1], we use the authors in the WWW conferences from 2001 to 2008 as persons and terms in their paper titles as attributes. Each person has $\sim$93 attributes and $\sim$7 links on average. Co-authorship prediction is considered as link recommendation problem for DBLP. For IMDB[2], we take all the actors/actresses who have performed in more than 10 movies since 2005 and consider movie locations as their attributes. There are 6750 persons and 9851 attributes. Each person has $\sim$29 attributes and $\sim$97 links on average. Co-staring prediction is considered as link recommendation problem for IMDB.

In order to test and compare different methods, we remove some edges in the graph and recommend the links based on the pruned graph. Four-fold cross validation is used. To test the results quantatively, we randomly sample 100 people and recommend the top-k links for each person. Precision, recall and mean reciprocal rank (MRR) are reported in Tables 1 and 2. LINKREC is compared with other methods based on the attribute and structure. SimAttr is to recommend links using attribute similarity. CommonNeighbors, Jaccard, Adamic/Adar and Katz are methods based on graph structure in [2]. SVM refers to Support Vector Machine on the combination of attribute and structure features. LINKREC(G), LINKREC(L) and LINKREC(M) denote the variants of our methods using global, local attribute weighting and mixed weighting schema. Our methods outperform all the baseline methods consistently in all the measures. Among the variants of our method, LINKREC(M) has the best precision, whereas the basic LINKREC performs the best in recall.

Table 1: Comparison of the methods on the DBLP data set

	P@1	P@5	P@10	P@20	Recall	MRR
SimAttr	0.3625	0.1455	0.0950	0.0603	0.5791	0.4478
CommonNeighbors	0.5775	0.2725	0.1708	0.1028	0.8155	0.6646
Jaccard	0.5625	0.2720	0.1708	0.1048	0.7998	0.6540
Adamic/Adar	0.6275	0.2985	0.1873	0.1094	0.8226	0.7127
Katz	0.5725	0.2675	0.1755	0.1045	0.8188	0.6641
SVM	0.6225	0.2985	0.1857	0.1099	0.8212	0.7068
LINKREC	0.7000	0.3470	0.2137	0.1228	**0.9068**	0.7767
LINKREC(G)	0.7350	0.3530	0.2175	**0.1239**	0.8912	0.7950
LINKREC(L)	0.7225	0.3345	0.1990	0.1135	0.8589	0.7882
LINKREC(M)	**0.7475**	**0.3605**	**0.2187**	0.1224	0.8809	**0.8058**

Table 2: Comparison of the methods on the IMDB data set

Method	P@1	P@5	P@10	P@20	Recall	MRR
SimAttr	0.6625	0.5355	0.4240	0.2914	0.3606	0.7384
CommonNeighbors	0.8475	0.7395	0.6525	0.5044	0.6390	0.8999
Jaccard	0.8775	0.7705	0.6835	0.5473	0.6694	0.9134
Adamic/Adar	0.8450	0.7570	0.6695	0.5184	0.6655	0.8992
Katz	0.7425	0.6710	0.5840	0.4446	0.5763	0.8332
SVM	0.8550	0.7590	0.6788	0.5531	0.7119	0.9002
LINKREC	0.8775	0.7660	0.6832	0.5544	**0.7243**	0.9172
LINKREC(G)	0.9100	0.7985	0.6935	0.5508	0.7234	0.9382
LINKREC(L)	0.9450	0.8140	0.7025	0.5429	0.6938	0.9609
LINKREC(M)	**0.9525**	**0.8180**	**0.7058**	**0.5591**	0.7230	**0.9648**

For the parameter setting in our framework, we obtain the best values by performing a grid search over ranges of values for these parameters and measuring accuracy on the validation set. λ controls the tradeoff between attribute and structural properties. Optimal λ is achieved at 0.6 in DBLP and 0.2 in IMDB, and the combination of attribute and structural features is much better than using attribute or structure properties individually. For the restart probability of random walks, it is interesting to find that a larger one is preferred and we set it as 0.9 in LINKREC. In other applications such as personalized search and query suggestion, random walks are used to discover relevant entities spread out in the entire graph, so a small restart probability is favorable in these cases. However, in link recommendation, we are more focused on the local neighborhood information, so a larger value is preferred.

5. CONCLUSION

In this paper, we proposed a framework for link recommendation based on attribute and structural properties in social networks. We presented several desired criteria for link recommendation and proposed a random walk algorithm on an augmented graph with both user attribute and graph structure to satisfy those criteria.

6. REFERENCES

[1] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explorations*, 7(2):3–12, 2005.

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

[3] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *ICDM*, pp. 613–622, 2006.

[1] http://www.informatik.uni-trier.de/$\sim$ley/db/

[2] http://www.imdb.com