A Survey on Link Prediction

Sourabh Vartak

M.S. Graduate Student
Department of Computer Science
State University of New York
Binghamton, NY – 13902, U.S.A.
svartak1@binghamton.edu

May 15, 2008

1. Introduction

Social networks are graph structures whose nodes or vertices represent people or other entities embedded in a social context, and whose edges represent interaction or collaboration between these entities [4]. Social networks are highly dynamic, evolving relationships among people or other entities. This dynamic property of social networks makes studying these graphs a challenging task. A lot of research has been done recently to study different properties of these networks. Such complex analysis of large, heterogeneous, multi-relational social networks has led to an interesting field of study known as Social Network Analysis (SNA).

Link prediction, the focus of my survey, is a sub-field of social network analysis. Link prediction is concerned with the problem of predicting the (future) existence of links among nodes in a social network [8]. Link prediction is the only sub-field of SNA which has focus on links between objects rather than objects themselves. This makes link prediction interesting and different from traditional data mining areas which focus on objects. In the figure below, link prediction lies in the circled area of SNA.

Link prediction is applicable to a wide variety of areas like bibliographic domain, molecular biology, criminal investigations and recommender systems. Most of the
surveyed papers use bibliographic domain data as it proves to be good representative data for link prediction. Also, such data is readily available on the Internet sites like DBLP, BIOBASE, CiteSeer, etc. Thus, most of the papers use co-authorship as a running example.

Link prediction problem can be posed as a binary classification problem that can be solved by employing effective features in a supervised learning framework [7]. Thus, in every paper that I surveyed, link prediction is done using topological features of the social network i.e. existing links, and various classification algorithms are applied to classify future link as Yes/No.

2. Framework

Link prediction addresses four different problems as shown in the figure below. Most of the research papers on link prediction, including which I surveyed, concentrate on problem of link existence (whether a [new] link between two nodes in a social network will exist in the future or not). This is because the link existence problem can be easily extended to the other two problems of link weight (links have different weights associated with them) and link cardinality (more than one link between same pair of nodes in a social network). The fourth problem of link type prediction is a bit different which gives different roles to relationship between two objects.

By doing survey of various research papers on link prediction, I have designed a framework in which the problem of link prediction is tackled as a binary classification problem. Classification whether a link exists or not can be performed using various supervised learning/classification algorithms like decision tree, K-nearest neighbors or support vector machines (SVM).

Different features like topological features, content/semantic information of individual nodes can be used for analyzing the proximity (relative closeness/similarity) of nodes in a social network. This feature information combined with different approaches like relational data approach and initial clustering approach help us predict the existence of a link between two nodes.
3. Discussion of different approaches and algorithms

The various papers which I surveyed have different approaches and algorithms to tackle the problem of link prediction. All these papers fit into the framework discussed in previous section in one way or the other. Referring to the framework, all the papers use the Topological features/Network structure, Apply proximity measure and Supervised Learning/Classification Algorithm blocks. This is so because it does not make any sense to predict future links without using current link information and treating nodes as isolated objects, thus, topological features must be used. Also, all the papers pose the link prediction problem as classification problem and thus use one or the other classification algorithm to predict links in social network. All the other blocks, namely, Relational data approach, Clustering used, Using content/semantic or individual node attributes, Joint/Conditional probabilistic model and Apply n-fold cross validation are optional depending on the specific approach taken by the paper.
Firstly, I will discuss the approach of performing initial clustering of nodes in the social network and then performing link prediction within these closely related nodes of same cluster. This approach is taken in the paper “Discovering Missing Links in Wikipedia”. In this paper, the authors have proposed a two step approach for discovering missing links in Wikipedia. The first step concerns identification of topically related pages, i.e., clustering. Identification of a cluster involves searching for certain graph structures. Two link-based similarity measures are co-citation and bibliographic coupling. In case of co-citation, two pages are related if they are co-cited by a third document. The strength of the relation is usually measured by the number of co-citation counts. In case of bibliographic coupling, two pages are related if both cite the same document. In this paper, authors have used co-citation idea. Authors have proposed LTRank (“Ranking based on Links and Titles”) algorithm which applies widely used similarity measure from information retrieval and uses authors’ own version of the Lucene full-text search engine. The second step involves identification of missing links. Search for missing links is confined to set of similar pages found in first step. Their hypothesis is that similar pages should have similar link structure. They identify candidate missing links and filter them through the anchor texts [1]. To recapitulate, approach of initial clustering tries to reduce the scope of graph in which we want to perform link prediction. This approach seems to be more scalable for very large graphs, where it will be unfeasible to perform link prediction on global scale.

The paper called “Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient” does not perform initial clustering of network objects but a topological feature called clustering coefficient is generalized to capture higher-order clustering tendencies. Clustering coefficient describes the tendency to form clusters (fully connected subgraphs) in a graph [2]. A typical clustering coefficient describes the probability for a connected triple to form triangles. Authors formalize the notion of generalized clustering coefficients to describe the formation of longer cycles. The proposed approach consists of a cycle formation link probability model, a procedure for estimating model parameters based on the generalized clustering coefficients, and model-based link prediction generation [2]. Thus, this paper does not perform clustering but uses a feature which measures clustering tendency of the nodes of the social network. It also uses the Joint/Conditional probabilistic model block of my framework but the model is not the Markov network (to be discussed in next paragraph). Also, the link prediction problem on abstract graphs (networks of no vertex and edge attributes) is the focus of this paper. Thus, the approach in this paper does not use ‘content/semantic or individual node attributes’ block of my framework.

The paper titled “Label and Link Prediction in Relational Data” addresses the task of collective label and link classification, where authors are simultaneously trying to predict and classify an entire set of labels and links in a link graph. This paper uses relational data approach as well as joint/conditional probabilistic model of Markov network. A Markov network, or Markov random field (MRF), is a model of the joint probability distribution of a set of random variables. Let V denote a set of discrete random variables and v an assignment of values to V. A Markov network for V defines a joint distribution over V. It consists of an undirected dependency graph, and a set of parameters associated
with the graph [5]. Since this paper uses relational data approach, authors have defined a relational schema for the Web hyperlink dataset they are using for their experiments. This schema consists of two entities: Page and Hyperlink. Attribute Page.HasWord indicates whether the word ‘k’ occurs on the page and Page.Label indicates topic of the page. Hyperlink.Type indicates type of relationship between two pages, Hyperlink.From and Hyperlink.To. A relational Markov network (RMN) specifies the cliques and potentials between attributes of related entities at a template level, so a single model provides a coherent distribution for any collection of instances from the schema. RMNs specify the cliques using relational clique templates to identify tuples of variables in the instantiation in a relational query language [5]. To recapitulate, this paper combines the relational data approach with joint/conditional probabilistic model of Markov network to perform the task of link prediction.

Now, I will discuss a paper which uses the relational data approach but does not combine it with probabilistic model as seen in above discussed paper. In this paper titled ‘Statistical Relational Learning for Link Prediction’, authors apply statistical relational learning to the problem of citation prediction in the domain of scientific publications. They formulate the feature generation process as search in the space of relational database queries. Aggregation or statistical operations, groupings, richer join conditions, or argmax-based queries can all be considered as part of search [6]. They have used the data from CiteSeer, an online digital library of computer science papers and designed a schema Citation (from:Document, to:Document), Author (doc:Document, auth:Person), PublishedIn (doc:Document, vn:Venue), WordCount (doc:Document, word:Word, cnt:Int). Citation provides topological or link information while WordCount provides content/semantic information. A query in relational algebra results in a table of all attribute values satisfying it, rather than a true/false value. Query results are aggregated using aggregate functions like count, avg, max, min, mode, and empty to produce scalar numeric values to be used as features in statistical learning. Aggregations can be applied to a whole table or to individual columns, as appropriate given type restrictions, e.g. avg cannot be applied to a column of a categorical type [6]. Adding aggregation operators results in a much richer search space [6]. The search space is potentially infinite, but not all subspaces will be equally useful. Authors propose the use of sampling from subspaces of the same type performed at the time of node expansions to decide if more thorough exploration of that subspace is promising, or if the search should be more quickly refocused on other subspaces. To summarize, this paper has proposed use of relational data approach to tackle the problem of link prediction by means of relational algebra queries including aggregate functions fired on a relational schema to predict future links. This paper does not perform initial clustering and does not use any probabilistic model like Markov network.

Now, I look at a paper which does not perform initial clustering, does not use relational data approach, and does not use probabilistic model. Thus, this paper titled ‘Link Prediction using Supervised Learning’ focuses only on different topological features, different aggregated features and various types of classification algorithms used to perform supervised learning. Essentially, this paper performs a comparative study of various topological and aggregated features applied to various classification algorithms to
predict links in a supervised learning setup. In this research paper, authors have used two bibliographic datasets: Elsevier BIOBASE and DBLP. Various topological features used are shortest distance, clustering index, Jaccard’s coefficient and shortest distance in Author-Keyword graph (extended social network by adding Keyword nodes) [7]. The clustering index measures the localized density. A node that is in dense locality is more likely to grow more edges compared to one that is located in a sparser neighborhood [7]. Various aggregated features used are sum of papers, sum of neighbors, sum of keyword counts, and sum of logarithm of secondary neighbors count. Since the number of secondary neighbors in social network usually grows exponentially, authors have taken the logarithm of the secondary neighbor count of the pair of nodes before they sum the individual node values. Various classification algorithms used are SVM (two different kernels), decision tree, multilayer perceptron, K-nearest neighbors (different variations of distance measure), naïve bayes classifier and bagging (groups the decisions from a number of classifiers). Authors have observed that SVM (especially with RBF kernel) performs the best for both the datasets BIOBASE and DBLP.

Finally, I will discuss a paper titled ‘Local Probabilistic Models for Link Prediction’ which neither performs clustering nor uses relational data approach but exploits all the three feature sets, namely, topological features, content/semantic information and probabilistic model. Firstly, co-occurrence probability features are derived by enumerating all simple paths of length K between two nodes. The frequency score of a path is defined as the sum of the occurrence counts of all nodes along the paths. Then, all paths are sorted by length and frequency score. All nodes along such paths which satisfy some threshold value are added to the central neighborhood set C for two nodes. Next we retrieve all non-derivable itemsets that lie entirely within this set. Their occurrence statistics are collected from the log events. We learn a local probabilistic model M over C. This local Markov Random Field (MRF) M specifies a joint distribution on all variables in C [8].

Katz measure is a weighted sum of the number of paths in the graph that connect two nodes, with shorter paths being given the more weight [8]. In this paper, modified Katz score which only considers paths of length up to certain threshold is used as the topological feature.

Semantic feature is derived by collecting the words in the titles of each author of co-authorship network. Then, derive a bag of words representation for each author, weighting each word by its TFIDF (Term Frequency - Inverse Document Frequency) measure. Then, compute the cosine between the TFIDF feature vectors of the two authors whose semantic similarity we need to determine. [8]

The three types of features - the co-occurrence probability feature, the topological similarity feature and the semantic similarity feature are combined using supervised learning framework. It was observed from the results that the co-occurrence probability feature captures information a large chunk of which is not captured by either topological metrics such as Katz or content-based metrics such as semantic similarity. Thus, the paper concludes that local probabilistic models like local MRF derived from co-occurrence probability are better at link prediction than topological or content-based features.
4. Conclusion

Link prediction is concerned with the problem of predicting the (future) existence of links among nodes in a social network or graph. The surveyed papers use various features like topological features, content/semantic information and probabilistic model like Markov network. Some papers use relational data approach and design a relational schema to better utilize link information, some others use initial clustering to localize the link prediction search space, and some papers combine various approaches. But, all the papers use link information and model link prediction problem as binary classification problem. Most of the link prediction techniques used in the research papers surveyed by me fit into the framework designed by me.

5. Acknowledgement

I would like to thank Prof. Lei Yu for his helpful discussions and comments throughout the semester on this survey regarding link prediction.

6. References

1. Discovering Missing Links in Wikipedia - Sisay Fissaha, Adafre Maarten de Rijke – LinkKDD 2005

2. Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient – Zan Huang - LinkKDD 2006


4. The Link Prediction Problem for Social Networks - David Liben-Nowell, Jon Kleinberg

5. Label and Link Prediction in Relational Data – B. Taskar, P. Abbeel, M. Wong, D. Koller


7. Link Prediction using Supervised Learning - M. Hasan, V. Chaoji, S. Salem, M. Zaki