A Survey on Link Prediction in Social Networks

Submitted by: Deepak Neralla

neralla2

Abstract

Social networks are highly dynamic objects which grow and change very quickly over time with the addition of new edges (connections, groups, relationships etc) signifying the appearance of new interactions and changes in underlying social structure.

One problem that is not well understood is the fundamental structure of how social networks evolve over time. i.e

*Given a snapshot of social network at time t , accurately predicting the connections interactions and relationships that will be added from the time t to a future time t’.*

Some of the applications of successfully modelling the above problem can be the analysis of a user's navigation history to generate tools that  increase navigational efficiency, accelerating  a mutually beneficial academic connection, overcoming data sparsity problem in recommender systems using collaborative filtering, identifying structure of criminal network etc.

Introduction

Social networks are a popular way to represent interactions within a group. These networks are dynamic in nature, which means that their configuration keeps changing with time, i.e. new nodes and connections are added frequently. A social network can be represented as a graph structure with a set of nodes and the links between them. The nodes represent people or other entities on a social network and the links are representative of a relationship, interaction or collaboration among these entities. Link prediction is a field under social network analysis. It deals with the problem of predicting links between the nodes of a social network graph. Apart from social networks, link prediction finds uses in several fields such as molecular biology, recommender systems in e-commerce and information retrieval. For example, in web domain, link prediction can be used for automatically hyperlinking web pages. Link prediction has also been used in the past for the prediction of protein-protein interaction in bioinformatics.

Broadly, three types of models are applied currently for solving the link prediction problem. These are:

1. Binary classification based (feature based) traditional supervised learning frameworks
2. Probabilistic models that work with joint-probability among graph entities, and
3. Linear algebraic approach for similarity computation between network nodes.

This survey briefly reviews the methods employed under each of the above models, and analyzes their strengths and weaknesses in terms of model complexity, prediction performance, scalability and generalizability. The survey concludes with a discussion on recent developments and the future of this field.

Notation

Consider a social network graph represented as G(V, E) in which an edge e = (u, v) ∈ E represents some form of interactions between two endpoints U and V at a particular time t(e). If we have multiple edges between two nodes, these can be represented by using a timestamp on each edge.

For a time t ≤ t′, we assume that G[t, t′] denotes a subgraph of G restricted to the the edges with time-stamps between t and t′.

1. Feature Based Link Prediction

Link prediction can be modeled as a supervised classification task where learning is done during a training period [t0,t0’] and testing phase is used to predict the links that will be added during a test period [t1,t1’].

Assuming that the interactions between two nodes u and v are symmetric, so that the pair <u,v> and <v,u> represent the same edge, the feature based link prediction can formally be stated as:

y<u,v> = +1 if <u,v> belongs to E

-1 if <u,v> doesn’t belong to E

where y<u,v> represents a label for the data point <u,v> (i.e the edge between u and v).  
This is a binary classification problem which can be solved by using any common supervised learning algorithm such as decision tree, naïve Bayes, kNN, SVM etc.

Feature Set

Choosing the appropriate feature set is the most important aspect of any machine learning algorithm. In the case of link prediction, these feature sets are largely dependent in the domain being analyzed. Features which represent some form of graph nodes proximity are chosen. In most works in this area, the features extracted from graph topology, and referred to as graph-topological features are used. Some of these features are:

1. Proximity Based Features

* Common neighborhood: Two nodes x and y are said to have a common neighborhood if they are both connected to the same node z. The probability of a link existing between x and y has been found proportional to the size of their common neighborhood, which means that if two nodes have a lot of common neighbor nodes, then there is a high chance of existence of a link between them as well.

In terms of authors of a journal[3], this can be gauged by the number of common keywords between the papers published by two different authors. The probability of two authors turning into collaborator is much higher if their papers have many common keywords.

* Jaccard coefficient can be used to normalize the size of common neighbours.
* Adamic/ Adar Measure: Adamic and Adar [4] have proposed a scoring function as:

1. Path Based Features

The smallest hop count distance between two nodes can be used to predict a future link between them. Intuitively, in terms of social network, we often become friends with our friends’ friends. Another variation of the path distance is based on the ensemble of all paths between two nodes x and y and was proposed by Katz [5]

where is the set of all paths of length l from x to y. βl is used for exponential damping.

Another path based feature is hitting time, which is defined as the average number of steps required for a random walk starting from x to reach y.

1. Clustering Index

It has been shown that a node with a dense neighborhood is more likely to develop edges over time compared to a sparse node.

1. Features Based on Vertex Attributes

Another kind of feature set is called aggregated features pertains to a single node. These are obtained by aggregating individual features which do not relate two nodes, but are a characteristic of a single node.

* Sum of Neighbors: The link prediction between nodes x and y takes into account the sum of number of neighbors of x and y. Accuracy can be improved by taking a weighted sum where the weight is based upon the number of common links with that particular neighbor. This measure derives from the clustering based measure described above.
* Sum of log(Secondary Neighbors Count): The number of secondary neighbors also sometimes play an important role. We take the logarithm since the number of secondary neighbors generally increases exponentially.
* Kernel Feature Conjunction: The basic idea is to obtain a kernel function that computes the similarity between two pairs of instances from the feature space which is expanded through Cartesian product.
* Extended Graph Formulation: We can extend a social network graph by adding additional nodes, corresponding to attribute features of existing nodes. Nodes sharing an attribute can be connected to this new attribute-node. The topology based metrics can now be applied on this new graph, so that both graph topology and node attributes are taken into account for similarity calculation.

Classification Models

Although there are a plethora of classification algorithms in the machine learning domain and most give comparable results, some are considered superior to the others for any particular domain.

The biggest challenge in supervised link prediction is the class skewness. The number of actual links in a graph is a very small fraction of the number of possible links, which affects the inference from training data. Moreover, this ratio further drops with time, which also needs to be accounted. For every new node added, the negative links grow quadratically whereas the positive links grow only linearly. For example, [6] reports that a period of 10 years, from 1995 to 2004 the number of authors in DBLP increased from 22 thousand to 286 thousand, thus the possible collaborations increased by a factor of 169, whereas the actual collaborations increased by only a factor of 21. To cope with this problem, authors in [7] suggest the alteration of training set by up-sampling or down-sampling. Other ways can be making the learning process cost sensitive, and taking variable thresholds for classification.

Another challenge specific to supervised link prediction is classification model calibration. Model calibration maps the output of a classification model to a particular class label and can impact the ratio of false positives and false negatives. The exact calibration is again domain dependent, determined by the cost associated with missing a true link or predicting a false link.

The training cost for supervised link prediction is also high in terms of time complexity and resources because if large sizes of social networks and may play a role in selection of classification model. Another important consideration is that the classification model should be dynamically updatable when new nodes and edges get added to the network.

Authors in [3] compared various classification models for co-athourship graph link predicition, The results for DBLP bibliography dataset are given in table 1.

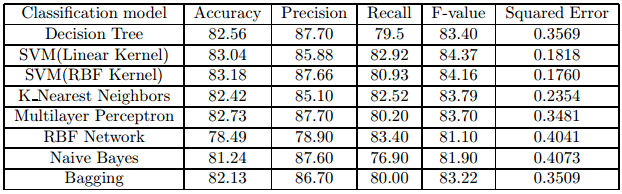


Table 1: Performance of different classification algorithms in DBLP dataset.

1. Probabilistic Link Prediction

*Locality Based Probabilistic Model*

Authors in [7] have presented a local probabilistic model that computes the joint probability of co-occurrence of nodes. According to the authors, the probabilistic model captures information not contained in the feature based topological model. This model can be used either as a standalone predictor, or with conjunction with topological models. This model has been found to be scalable and highly discriminatory.

The first step in this process is to identify a central neighborhood for two nodes x and y consisting of nodes that are neighbors of either of these nodes. This is determined by the local graph topology. We use a local model, because constructing global models is very complex and expensive. Once the central neighborhood, say (w,x,y,z), is identified, a Markov random field(MRF) model is learnt to estimate the joint probability of w,x,y and z.

We can include all nodes lying in the path between two given nodes in their central neighborhood but this causes the model size to be prohibitively large, so we define a parameter ‘size’ to denote the number of nodes in central neighborhood. Intuitively, the nodes lying on shorter length paths are more valuable hence we start accumulating neighborhood nodes starting from shorter paths and expanding till we reach ‘size’ value. We also define a ‘frequency score’ for tie-breaking in case of equal length paths. This is defined as the sum of the occurrence counts of all nodes along a path. Figure 1 gives the pseudo code of this central neighborhood calculation algorithm.

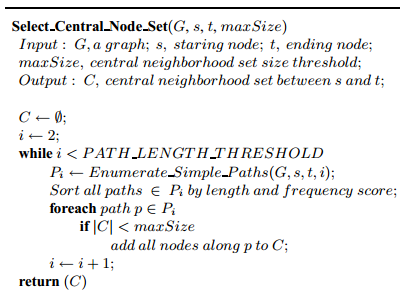


Figure 1: Selecting the central neighborhood set for two nodes [7]

To learn the local Markov random fields once we have a pair f nodes and their central neighborhood, the authors have proposed an approach involving non-derivative frequent itemsets obtained from the underlying event log of graphs. The frequent itemsets mining techniques can be applied in this scenario and the frequent itemsets along with their occurrence statistics are used to formulate a constraint for the underlying unknown probability distribution to be discovered. The model that satisfies all these constraints and has maximum entropy (indicating a uniform distribution) is selected. Since social networks are generally sparse, i.e. the number of existing links are a small fractuion of possible links, we use a support threshold based mining algorithm as describes in [8].We use non-derivable frequent itemsets so that the constraints provided by them are not redundant, as these itemsets cannot be derived from other itemsets. The implementation provided in [7] precomputes all such itemsets and then selects the frequent items sets pertaining to the nodes present in the central neighbourhood and uses them to learn the MRF. The formal algorithm for this is given in figure 2.

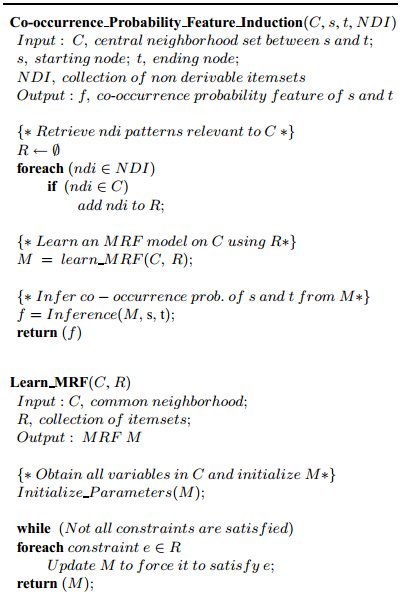


Figure 2: Inducing co-occurrence probability feature for a pair of nodes.

*Network Evolution Based Probabilistic Model*

This is another model based on graph topology, but it focuses on the network topology rather than the information provided by the nodes themselves. This paper presents a novel parameterized edge label copying model for network structure evolution which can be further used for link prediction. The structure of network is assumed to probabilistically change over time. This is given by an edge label function φ which indicates whether an edge exists between two given nodes. Since this may vary over time, we add a time component and represent it as φ(t). The vertices V are fixed through time. Given a network state {V, φ(t)} at a time t, the problem then reduces to one of probabilistically modeling any flips in the values of φ. We assume a Markov model so that the state at time t+1 depends only on the state at time t and perform a “copy-pasting” of edge labels governed by certain network criteria.

Assume that an edge label is copied from a node l to node m randomly with a probability wlm. This, node l randomly selects one of its edges to other nodes and copies the label to its edge to node m. This is controlled by the following probability constraints:

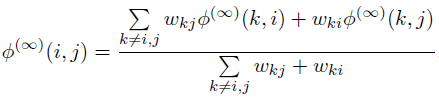
, and

Also, wll = 0 which means that copying of an edge from a node to itself is not permitted. Intuitively, if a node k has a strong influence on node i, then the edges from node k will likely be duplicated into node i. Conversely, an absence of edge from k to another node j will likely indicate no edge between i and j as well. If thought in the field of biotechnology, this can be seen as a copied gene in biological evolution having the same characteristic as its parent (influencer) gene.

An EM based transductive learning approach is applied. The probability of an edge existing between two nodes i and j is then the sum of probability of an edge being added between i and j at time t+1, and the probability of an edge already existing between i and j at time t and remaining unchanged over time. This can be written as:

|V|-1 represents all possible edge labels from node k. The first term indicates that the probability was unchanged therefore, at time t, there were already edges (k,j) and (i,j) in existence. The network structure evolves over time by iterative applications of this equation.

For the task of link prediction, we assume a stationary state of network evolution, which means that φ(t+1) = φ∞ = φ(t). Substituting this in the above equation gives the new value for φ(t+1) as:



The estimation problem can then be formulized as maximization of φ∞ (i,j) L(W)

where L(W) is the log likelihood of known edge labels and is defined as:

El is the set of all known edge labels. Figure 3 summarizes this algorithm:

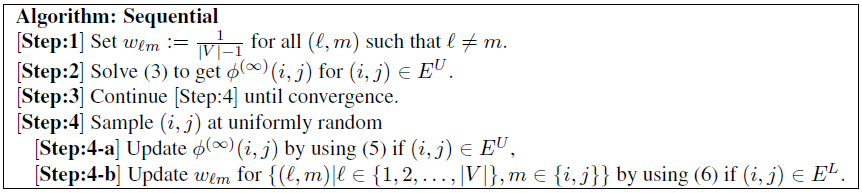


Figure 3: A sequential transduction algorithm for link prediction [8]

References

[1] A Survey on Link Prediction, Sourabh Vartak May 2008

[2]LINK PREDICTION IN SOCIAL NETWORKS, Mohammad Al Hasan, eBay Research Labs, Mohammed J. Zaki

[3] Link Prediction using Supervised Learning, Mohammad Al Hasan Vineet Chaoji Saeed Salem Mohammed Zaki

[4] Adamic, Lada A. and Adar, Eytan. (2003). Friends and neighbors on the web. Social Networks, 25(3):211-230.

[5] Katz, Leo. (1953) *A new status index derived from sociometric analysis*. Psychometrika, 18(1):39-43.

[6] Rattigan, Matthew J., and Jensen, David. (2005). *The case for*

*anomalous link discovery*. SIGKDD Explorations Newsletter.

[7] Local Probabilistic Models for Link Prediction, Chao Wang, Venu Satuluri, Srinivasan Parthasarathy

[8] J. Han, J. Pei, Y. Lin. Mining Frequent Patterns without Candidate Generation, ACM SIGMOD International Conference on Management of Data.

[9] A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction, Hisashi Kashima, Naoki Abe, IBM Research