RESEARCH ARTICLE OPEN ACCESS

An Enhanced Supervised Random Walk Algorithm for Link Prediction

A. vihashini [1], Dr. G.T. Prabhavathi [2]
Research Scholar [1], Associate Professor [2]
Department of Computer Science
Gobi Arts & Science College, Gobichettipalayam
Tmil Nadu - India

ABSTRACT
The problem of link prediction is to predict the links that does not exists and have the probability to occur in the near future. A random walk is a mathematical formalization of a path that consists of a succession of random steps. Node and link attributes along with node structure information are used for prediction. Supervised random walks are trained on a single snapshot of a graph at time t0 and immediately tests on its predictions for new links from a source node s at time t1. This paper presents an enhanced new link representation model based on implicit user feedback obtained from search engine using K-SVM queries. The main objective this approach is to achieve better results in supervised tasks, such as clustering and labelling, through the incorporation of usage data obtained from search engine queries. This type of model allows a user to discover the motivations of users when visiting a certain topic.

Keywords:- Link Prediction, Random Walks, Social Networks, SVM

I. INTRODUCTION

Social networks are extremely dynamic for developing relationships among people or other entities, they grow and change quickly over time through the addition of new edges, expressing the appearance of new interactions in the underlying social structure. Much of the research in mining social networks is focused on using the links in order to derive interesting information about the social network such as the underlying communities, or labeling the nodes with class labels. In the most social networking applications, the links are dynamic, and may change considerably over time. For example, in a social network, friendship links continuously formed/removed time by time. Therefore, an interesting question in the research of social network applications is to determine or predict future links that may occur in the network structure[4].The prediction process may use either the structure of the network or the attribute-information at the different nodes. Link predictions can be applied to predict how a specific product can be marketed in the future, to identify the formation of terrorist's network, spreading of disease in a geographical area, etc. A variety of structural and relational models have been proposed in the literature for link prediction [2].

Given a snapshot of a social network at time *t*, the problem of link prediction is to infer which interacting members are likely to occur in the future or which among existing interactions are missing during the interval from time t to a given future time t+1[1]. The links can be predicted using supervised or unsupervised approaches. Unsupervised methods assign a score for each pair of nodes with base on neighborhood nodes (local) or path (global) information. Global methods usually achieve higher accuracy than local methods. Unsupervised strategy considers the link prediction problem as a classification problem. Considering a social network, the network structure can be represented as feature vector for each pair of nodes. These vectors are used to train different classifiers to determine whether the link exist or not between a pair of nodes [3].

In supervised link prediction approach, the supervised learning strength is assigned to the edges that are likely to have new links. The strength is not set manually, but learned from the features of each edge and nodes between them. Random walk is a popular approach in supervised link prediction. Supervised random walk algorithms are applied in problems like graph recommendation, anomaly detection, expertise search and ranking and widely used in missing link prediction. Support vector machines are supervised learning models used to analyze data and recognize patterns. This paper describes SVM in chapter 2 and the dataset used for experimental evaluation in chapter 3.chapter 4 and 5 describes the experimental setup and the results of K-SVM algorithm.

Given a two-class separable training data set, intuitively a decision boundary is drawn in the middle of the data distribution of the two classes. The SVM in particular defines the criterion to be looking for a decision surface that is maximally far away from any data point. This distance from the decision surface to the closest data point determines the margin of the classifier[5].

II. SUPERVISED RANDOM WALKS

A Random walk is a stochastic process that consists of a sequence of discrete steps of fixed length. Given a graph and a starting point, a neighbor of it is selected at random and moved to this neighbor, and then neighbor of this point is selected at random, and moved to it, etc[6]. The (random) sequence of points selected this way is a random walk on the graph. Random walk algorithms are more suitable for link prediction problems. Random-walk based algorithms are based on the structure of the network and the nodes attributes. Node and link attributes along with node structure information are used for prediction [8].

Both supervised and un-supervised random walk models can be used for predicting the links in a social network. Supervised random walks are widely used for predicting links in the network as it demonstrates a good generalization and overall performance. Supervised learning strength is assigned to the edges that are likely to have new links. The strength is not set manually, but learned from the features of each edge and nodes between them. The primary limitation of a supervised approach is the reduced richness of the data representation.

To address the challenges in link prediction many algorithm have been developed for both link prediction and link recommendation. In the proposed model, a Supervised K-SVM is combined with the characteristics (attributes, features) of link prediction and edges of the network into a unified link prediction algorithm. In this method based on Supervised K-SVM, the algorithm learns how to bias a like random walk on the network it visits given topic link prediction (i.e., positive training examples) more often than the others[7]. It uses the node and edge features to learn edge strengths (i.e., random walk transition probabilities) such that the random walk on a weighted network and is more likely to visit “positive” than “negative” link prediction. The supervised K-SVM algorithm has some attractive properties. They can be easily estimated statistically.

The goal is to learn a function that assigns strength (i.e., random walk transition probability) to each edge so that when computing the random walk scores in such a weighted network s , it creates new links that have higher scores to s than link prediction to which it does not create links. From a technical perspective, that such edge strength function can be learned directly and efficiently.

The experiment with large collaboration networks and data from the browsed social network, shows that the proposed systems approach outperforms unsupervised approaches as well as supervised approaches based on complex network feature extraction. An additional benefit of this approach is that no complex network feature extraction or domain expertise are necessary as this algorithm combines the node attribute and network structure information.

This works presents an enhanced link representation model based on implicit user feedback obtained from search engine K-SVM queries. The main objective of this model is to achieve better results in supervised tasks, such as clustering and labeling, through the incorporation of usage data obtained from search engine queries. This type of model allows a user to discover the motivations of users when visiting a certain topic. The terms used in queries can provide a better choice of features, from the user’s point of view, for summarizing the web pages that were clicked from these queries.

K-SVM ALGORITHM

1) Set Start state to s0
2) Mark start state s0 as already visited s' = s0
3) While the length of data result not reached or exit criteria not reached repeat the following
   For all unvisited states k:
   - Compute P(s' --> s) using the support vector
   - Choose the Max(P(s'--> s)) and let the corresponding set of
   - states be S
   - If |S| > 1 then pick the state s" from S
   - if |S| = 0 then restart (for example: s" = s0)
   - s' = s"
4) Data result Generated

III. DATA SET

The input data to this algorithm is a set of DBLP URLs with their similarity information. For testing the system, 50

DBLP URLs from a set of 360 randomly chosen hosts with at least 250 journal links were crawled and standard K-SVM classification techniques were applied. In the proposed system, DBLP URLs that are in duplicate clusters of size at least 2 is kept.

Then, a 15-20 test-train split was performed by randomly assigning each cluster of duplicate DBLP URLs to either the training set or to the test set. The use of two datasets: BIOBASE and DBLP, has information about different re-search publications in the field of biology and computer science, respectively. For BIOBASE, it used 5 years of dataset from 1998 to 2002, where the first 4 years are used as training sets and the last for testing. For DBLP, it used 15 years of dataset, from 1990 to 2004. First 11 years were used as training set and the last 4 years as test. Pairs of authors that represent positive class or negative class were chosen randomly from the list of pairs that qualify. It constructed the feature vector for each pair of authors. A detailed description of the features is given in the following sub-section. The datasets have been summarized in Table 1.

Dataset	Number of papers	Number of authors
BIOBASE	831478	156561
DBLP	540459	156561

Table 1: Statistics of datasets

IV. EXPERIMENTAL SETUP

(i) Content Analysis

Phrase finding, which enables the user to identify key concepts by browsing a list of automatically extracted phrases, is a useful tool for link prediction. There are three types of Data-oriented phrase finding capabilities in the system:

1. Given a set of messages, find key phrases which are commonly mentioned in the messages.
2. Given two sets of messages, find the set of key phrases that best discriminate the two sets.
3. Given a phrase and surrounding context from a set of messages, find collocations (words or phrases which frequently appear together with the specified phrase). In this model all the three content analysis were performed on the data before evaluating the data.

(ii) Evaluation

The objective of this algorithm is to evaluate the performance of implicit query prediction using K-SVM model in a series of controlled experiments. User click tracking data was collected from test subjects who read short text snippets trying to identify if they were about a given topic. Each test subject completed several sessions. The testing was done in a leave-one-session-out fashion: document relevance's on one session were hidden, and an implicit query was estimated using the document relevance and user click movement data from the other sessions. Documents on the test session were ranked according to the implicit query, and the resulting ordering was compared to a known ground truth. The model learned using the user click movement and textual features performed better than a similar model that used only textual features.

The evaluation is indeed to test the existence of a dependency between user click movements and interest, confirming the assumption. Furthermore, it is possible to utilize this dependency to do proactive retrieval by implicit user click movement feedback. An interesting question is which user click movement features help in discriminating between relevant and non-relevant words. For evaluating the results, in this model a analyzes based on regression coefficients of the linear model were performed and found three coefficients that differ significantly from zero: saccade length before first fixation to the word, indicator variable for a regression being initiated from the next word and duration of the first fixation to a word relative to the total fixation duration.

Both how long a word is being viewed (relative to other words) and the styles of saccadic movement between the neighboring words are indicative of the relevance of the word. The features used in this study have been initially proposed in various studies, not necessarily related to information retrieval. It is likely that it would be possible to construct features that are even better suited for the relevance prediction task.

V. RESULTS

Table 2 and 3 shows the performance comparison for different classifiers on the BIOBASE and DBLP datasets respectively. In both datasets, counts of positive class and the negative class were almost the same. The baseline classifier would have an accuracy around 50% by classifying all the test data point to be equal to 1 or 0, whereas all the models that were tried in this experiment has reached to an accuracy

above 80%. This indicates that the features that were selected have good discriminating ability. For BIOBASE dataset 9 features and for the DBLP dataset only 4 features were used.

Classification model	*Accuracy*	*Precision*	*Recall*	*F-value*
Decision Tree	*90.01*	*91.60*	*89.10*	*90.40*
SVM(Linear Kernel)	*87.78*	*92.80*	*83.18*	*86.82*
K_Nearest Neighbors	*88.17*	*92.26*	*83.63*	*87.73*
RBF Network	*83.31*	*94.90*	*72.10*	*81.90*
Naive Bayes	*83.32*	*95.10*	*71.90*	*81.90*
K-SVM	***90.56***	***92.43***	***88.66***	***90.51***

Table 2: Performance of different classification algorithms for BIOBASE database

Classifion model	*Accuracy*	*Precision*	*Recall*	*F-value*
Decision Tree	*82.56*	*87.70*	*79.5*	*83.40*
SVM(Linear Kernel)	*83.04*	*85.88*	*82.92*	*84.37*
K_Nearest Neighbors	*82.42*	*85.10*	*82.52*	*83.79*
RBF Network	*78.49*	*78.90*	*83.40*	*81.10*
Naive Bayes	*81.24*	*87.60*	*76.90*	*81.90*
K-SVM	***83.18***	***87.66***	***80.93***	***84.16***

Table 3: Performance of different classification algorithms for DBLP dataset

On accuracy metrics, K-SVM has performed better for both the datasets with an accuracy of 90.56% and 83.18%. Naturally, the performance on DBLP dataset is not as perfect when compared to BIOBASE as fewer features were used in the former dataset. Moreover, DBLP dataset was obtained using 15 years of published articles and the accuracy of link prediction deteriorates over the longer range of time span since the institution affiliations, coauthors and research areas of researchers may vary over the time. So, predicting links in this dataset is comparably more difficult than the BIOBASE dataset, only 5 years of data were used. In both the dataset, other popular classifiers like Decision Tree, KNN and SVM also have similar performances usually 0.5% to 1% less accurate than K-SVM. From the results, K-SVM proved to be slightly accurate than other compared measures.

The performance analysis of the proposed algorithm for the datasets BIOBASE and DBLP are shown in Figure 5 & Figure 6 respectively.

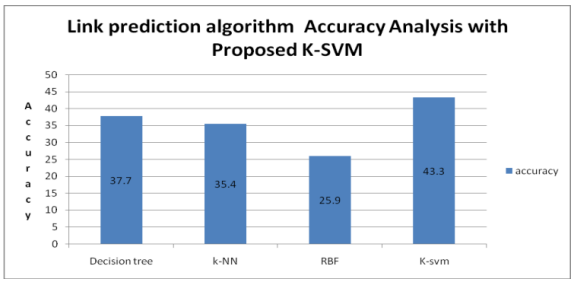

FIGURE 5: Accuracy Analysis with proposed K-SVM

VI. CONCLUSION

Data mining refers to extracting or mining of useful information from large amounts of records or data. Data mining includes the task of data clustering, association analysis and evolution analysis. For link prediction, one should choose features that represent some form of proximity between the pair of vertices that represent a data point. The definition of such features may vary from domain to domain for link prediction. In this research, co-authorship network has been taken. Random walk is one of the popular approach for predicting links in social networks. From the literature review, supervised random walks have been found to produce better results than unsupervised random walks. Hence, in this work, supervised random walk approach has been taken to solve the problem of link prediction.

SVM (Support vector machines) had been the most developed method for classification and regression technique as it improves the search performance and the user experience. The experiments show that it is possible to infer the interest from user click movements. The experiment with K-SVM user search complexity reduces by more than 90% the number of features needed to represent a set of users and improves by over 92% the quality of the SVM, KNN results.

This model significantly produces better features and provides more accurate labels according to the user's expectations.

REFERENCES

[1] Archana.S, and K.Elangovan, Survey of Classification Techniques in Data Mining, International Journal of Computer Science and Mobile Applications, Vol.2, Issue.2, pp. 65- 71, 2014.

[2] Backstrom and J. Leskovec, Supervised random walks: Predicting and Recommending links in social networks, WSDM '11 in Proceedings of the fourth ACM international conference, pp.635-644, 2011.

[3] Chakrabarti.D, R. Kumar, A.Tomkins. Evolutionary Clustering, ACM KDD Conference, 2000.

[4] David Liben-Nowell and Jon Kleinberg, The Link Prediction Problem for social networks, In Proceedings of the twelfth international conference on Information and knowledge management,NewYork,CIKM '03, pp. 556–559, 2003.

[5] Hsu, Chih-Wei; Chang, Chih-Chung; and Lin, Chih-Jen , A Practical Guide to Support Vector Classification, PDF, Technical report. Department of Computer Science and Information Engineering, National Taiwan University,2003.

[6] Hasan.M.A, V. Chaoji, S. Salem, and M. Zaki-Link Prediction using Supervised Learning, SDM 06 workshop on Link Analysis, Counterterrorism and Security, 2006.

[7] HaveliwalaT.H, Topic-Sensitive PageRank. In Proceedings of WWW'02, New York,USA, ACM, 2002.

[8] László Lovász- Random Walks on Graphs: A Survey, Paul Erd˝os is Eighty(Volume2)Keszthely (Hungary), pp.1–46, 1993.