Harnessing Social Network with Link Data Mining Approach for Predictive Analytics

Raghwendra Singh[1], Dr. Arjit Tomar[2], Kunwar Babar Ali[3], Jayati Mukherjee [4], Sneha Mishra[5]

[1] *School of Engineering, Department of Computer Science, Noida International University, Greater Noida, UP, India*
Email-rschandan.singh@gmail.com

[2]*Assistant professor, Department of Computer Science, Noida International University, Greater Noida, UP., India*
Email.-arjit.tomar@gmail.com

[3]*Assistant professor, Department of Computer Science, Noida International University, Greater Noida, UP., India*
Email.-kunwarbabarali1@gmail.com

[4]*Assistant professor, Department of Computer Science, Noida International University, Greater Noida, UP. , India*
Email.-kunwarbabarali1@gmail.com

[5]*Assistant professor, Department of Computer Science, Noida International University, Greater Noida, UP. India*
Email.-kunwarbabarali1@gmail.com

Abstract

Data bases, artificial intelligence, machine learning, social networks are many such areas where data mining leaves a great impact. They are playing important role in carrying out significant researches in the field of data mining. In today's world where data is growing at a very strong rate and information retrieval is becoming a very complex task, users are demanding more and more useful and significant information hidden in respected data. There are also various problems faced by many developers regarding data such as data incompleteness, data inconsistency and data missing because the data used for different motives of data mining applications may or may not be useful because of these problems. In today's world, one of the interesting fields that contributed significantly in data mining researches and provide a new era to data mining work is Social Network. Social network is a concept that is becoming very popular among individuals as it solves various purposes of their interests. The social network can be interpreted as "a structure of nodes and ties or edges." Here nodes represent the various objects (living or non living) and ties represent various relationships among them.

Keyword : *Data Mining, Data Analytics, Social Network Link Data Mining, Predictive Analytics.*

Introduction

Link Mining is that part of data mining that studies social network [6]. Link mining or link analysis can be seen as newly uprising area in the field of data mining that mainly focused on links between objects in spite of objects itself. It considers only relationships of individuals existed in respected social networks and extract out useful invisible patterns. Social network analysis is posing many severe challenges in terms of the appropriate methodologies, algorithms and effective implementations. The problem gets more complicated because of data capturing, data analytics and extracting useful pattern from the data. Many attempts for social network analysis are not so beneficial

because they have a need in terms of appropriate approach, direction, objective and suitable methodology. These challenges motivated me to do my research work in the field of Social network analysis [9].

Link prediction problem can be viewed as a simple binary classification problem [7]: For any two potentially linked objects Oi and Oj, predict whether Lij is 0 or 1.There are number of researchers suggested various efficient methods for link prediction. Some of the methods suggested [5][6] considered topological features and attributives features with temporal aspects to predict network structure but they have not considered normalization aspects of these features so that efficiency of predictor can further achieved. Now problem is to build a link predictor using these three features (topological features, semantic features, temporal features) at the same time but in a normalized way so as to achieve better performance from previous methods.

Methodology

We have considered the topological pattern based method as our link prediction approach while incorporating Jaccard's coefficient as the normalized measure of common neighbors [4]. As our approach is described in the figure shown below which provides systematic approach of link prediction in a social network.

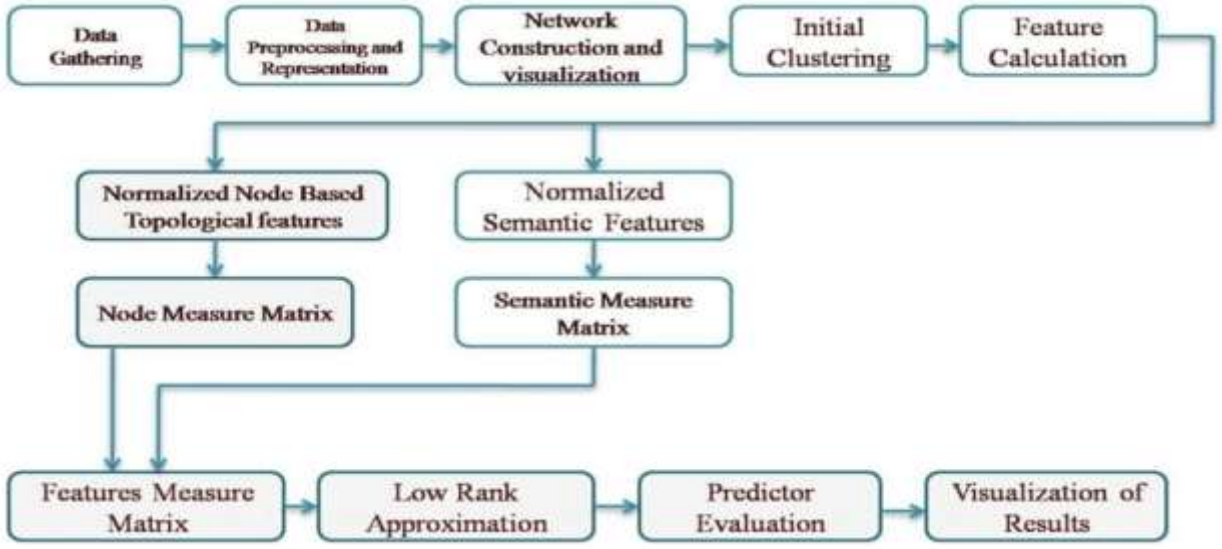

Fig.1. Methodology for Link Data Mining

Network Construction and Visualization

When data is preprocessed and converted into a suitable format then next step is to construct network. In constructed network bibliographic data is used to construct co authorship network in which two researchers are connected if they coauthored a paper. One co authorship network constructed from bibliographic information taken from DBLP website is shown in figure below. This network is visualized using a GEPHI tool which is an open source network visualization tool. It takes the data in .net format.

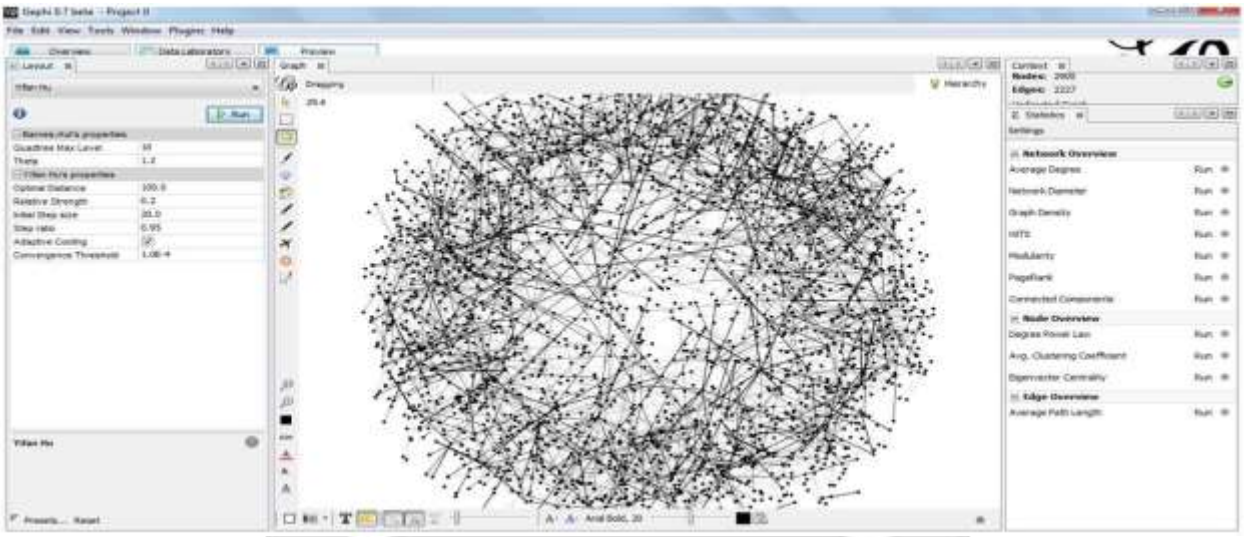

Fig.2. Co Authorship Network

This tool also provides some basics metrics and statistics which provide useful information about network. These metrics includes average degree of network, graph distance report including between's measure, closeness measure that is shown in figure below.

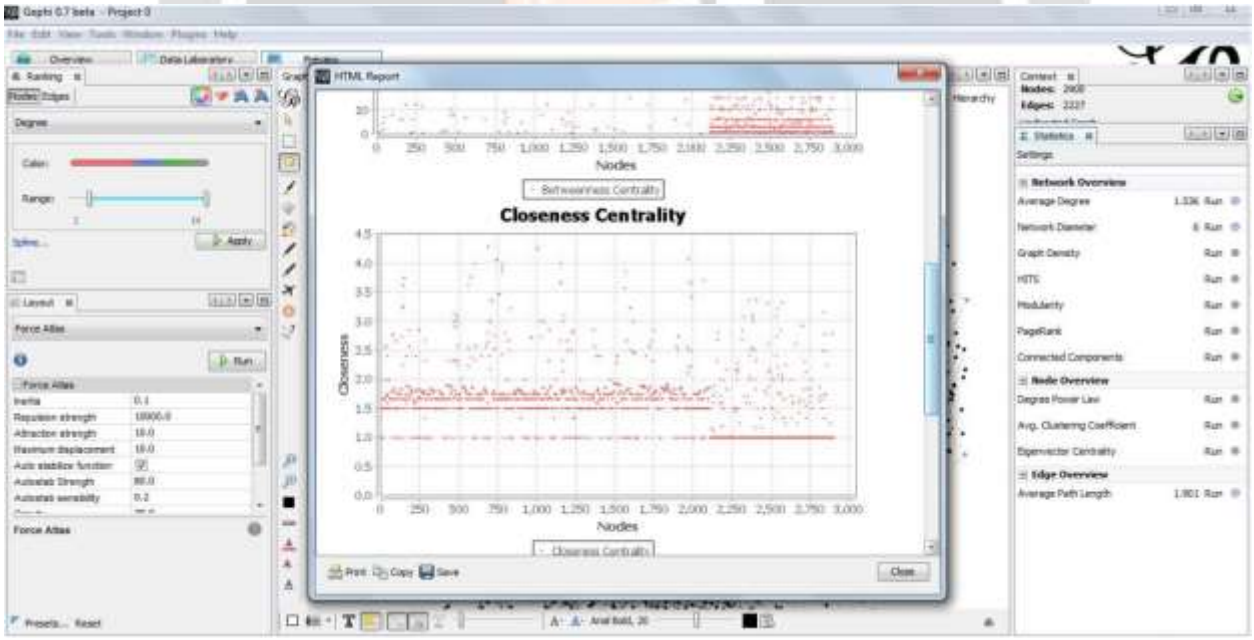

Fig.3. Initial result: Closeness Centrality

Node Based Topological Features

As name suggest, these features provides the structural view of given social network from the perspective of nodes of that network. These features take only information into consideration which is related to nodes of the network. The nodes of the network follow some common properties on the basis of which nodes are connected. These node based topological properties are given below-

Common Neighbors

One of the common properties exhibited by nodes is the number of common neighbors that each pair of node in the network has in common. This property is commonly used in collaboration network which confirms the quantity of common neighbors of x and y which contribute to the probability that they will collaborate in future.

Suppose Vi and Vj are two nodes in a given network then common neighbors of these two nodes can be computed as

$$\textbf{Score } (V_i, V_j) = |V_i \cap V_j|$$

Jaccard's Cofficient

This is a one of the metrics that is used for similarity measure in information retrieval measures. Here a feature is selected for similarity measurement between selected two nodes and then provides normalized version of it. For e.g. in a co-authorship network, if we consider common neighbor as Network feature then normalized score of common neighbors is given by

$$\textbf{Normalized Score (Vi,Vj)} = \frac{|V_i \cap V_j|}{|V_i \cup V_j|}$$

Semantic Features

Topological features are not sufficient to provide full information of a network used for making future predictions. So some additional information is also required that contributes in extraction of hidden information of network. This information can be taken as semantic information that varies network to network. This information can be considered as semantic features of a network.

Semantic information is nothing but the information that provides an insight to the network. The captured semantic information is then used to calculate the normalized score to compute the similarity between any two pairs of individuals in the corresponding network [4].

For e.g. in a co-authorship network of authors, the titles, abstract etc of the paper can be viewed as the semantic information provided by this kind of the network. We can extract useful information hidden in these types of data.

For e.g. if we consider titles of papers as the semantic information for co-authorship network then Normalized score given by,

$$\textbf{Normalized score (W,W)} = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}$$

Predictor Evaluation and Visualization of Results

This can be considered as the last step of methodology. Some metrics are also suggested by various other researchers to explain and judge the performance of predictor named as Precision, Recall and F-score used by [4]. They suggested that for any feature vector, predictor P can make either positive or negative prediction. In the positive case, if prediction is true then it is said to be true positive (TP), else it is false positive (FP). Likewise in negative case, prediction can be either false positive (FP) or false negative (FN). Now on the basis of these values, they defined the three performance metrics. They defined recall as how efficiently predictor can predict the future collaborations and can be defined as

$$\textbf{Recall} = \frac{|TP|}{|TP| + |FN|}$$

Precision gives the proportion of positive predictions that are true out of total positive future predictions and can be defined as,

$$\frac{|TP|}{|TP| + |FN|}$$

$$\textbf{Precision} = \frac{\qquad\qquad}{\qquad\qquad}$$

On the basis of these two metrics, final score of predictor is computed named as f-score which is the harmonic mean of recall and precision and can be defined as,

$$\textbf{F-Score} = \frac{\textbf{2*Precision*Recall}}{\textbf{Precision +Recall}}$$

With the help of these metrics, final values of predictor can be computed which are the probabilities of collaboration are between authors in near future. These values are computed between every pair of authors presented in given database. With these values a graph is plotted between predictor probabilities and author pair count which graphically shows the likely collaborations of authors in near future. Results are also visualized with the help of clusters of authors. Clusters are made on the basis of probabilities on which authors are likely to collaborate. These probabilities are divided into three levels e.g. likely, more likely and most likely on which clusters of authors are made.

Snapshot for clustered network and graph

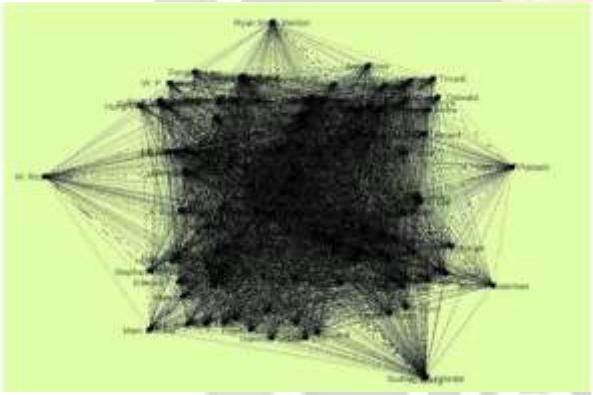

Fig.4. (a) Snapshot of clustered network

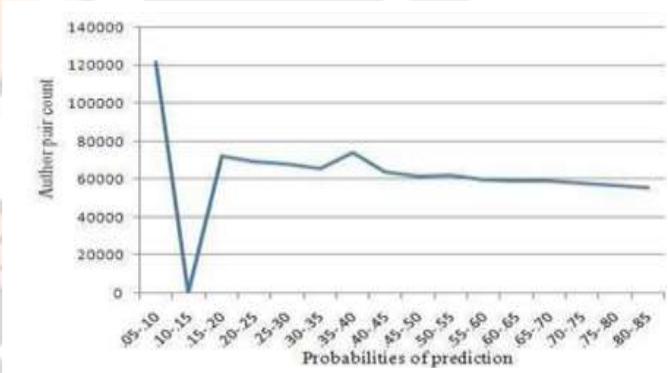

Fig. 4. (b) Graph of result

Conclusions

This thesis work extends the work done previously in the field of link prediction in social networks. Our research work proposed a framework that extends the existing work by including various other network features that were not included before and then matrix based approach is applied to gain the fruitful results. The methodology of social network analysis proposed by us provides a step by step approach from network construction to feature creation calculation and leads to link prediction in social networks. The described approach to predict the collaboration aims to predict the probable collaboration among researchers of common interests that are the part of co- authorship network based on various network parameters that are proposed in this research work. It is believed that collaborations will hold the key for survival in competitive world. In our proposed study of social network analysis we have considered the academic social network for research collaboration. Our proposed methodology framework forms the foundation for predicting future collaborations and considers significant aspect of social network being diverse and having different features. Some of the existing methods have though explored these features but have not considered a significant aspect which can be termed as normalization. These various features can be well represented in our proposed feature measure matrix which includes attributive feature matrix and topological feature matrix with temporal factor which also considers normalization aspect of these features. The normalization aspect to

these features is required for enhancing the efficiency of predictor. The proposed methodology will yield insightful results in predicting future collaborations between various researchers and predict future links with greater efficiency.

Future Work

In this area of data mining, there is lot of future research work. As it is already said above, collaboration is the key concept in today's competitive world. Collaboration is way by which two or more people can share their ideas, knowledge to bring out some potential results. This work can be extended to other social networks such as friendship networks, biological networks, business networks etc. to extract the hidden information from different types of collaborations existed in different networks.

Different networks have different characteristics according to their structure so the different attributes that are considered in our proposed methodology can be extended with respected to given network as different networks have different semantic characteristics associated with them on the basis of which we can decide semantic attributes for those networks.

This work is also closely related to one of the interesting field e.g. linked data where data are collected from different sources and then data is preprocessed and refined for different data mining purposes. So this concept can be extended in link prediction where different networks of same purposes and nature can be merged on the basis of some common relationships and then predictive analytics can be performed for extraction of hidden relationships more efficiently.

References

1. Loyalty Square: Social Network Analysis https://www.indiamart.com/loyaltysquare/aboutus.html".

2. Liben-Nowell, D., Kleinberg, J.: "The link prediction problem for social Networks": Twelfth international conference on Information and Knowledge Management, ACM Press, New York, pp. 556-559(2003).

3. Pavlov, M., Ichise, R.: "Finding experts by link prediction in co-authorship networks": 2nd International Workshop on Finding Experts on the Web with Semantics, pp. 42-55(2007).

4. Mrinmaya Sachan, Ryutaro Ichise. "Using Abstract information and Community Alignment Information for Link Prediction": Second International Conference on Machine Learning and Computing (ICMLC), ISBN: 978-1-4244-6006-9, pp. 61-65(2009).

5. Jingfeng Guo, hongwei Guo."Mullti-features Link Prediction Based on Matrix": International Conference on Computer Design and Applications (ICCDA) ISBN: 978-1-4244-7164-5, V1-357 - V1-361(2010).

6. Victor Stroele, Jonice Oliveria et. Al."Mining and Analyzing MultiRelational Social Networks": International Conference on Computational Science and Engineering, Print ISBN: 978- 1-4244-5334-4, pp. 711-716(2009).

7. Getoor, L., Diehl, C.P.: "Link Mining: A Survey ": ACM-SIGKDD Explorations, Volume 7, Issue 2(2005).

8. Evan Wei Xiang, Qiang Yang.: "A Survey on Link prediction Model for Social Network Data". Science and Technology (2008).

9. Tarun Kumar, O. P. Vyas.: "Harnessing Social Network with link Data Mining for Predictive Analytics: An Extended Approach": 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence (ICADABAI) (2011).

10. Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze.: "Introduction to Information Retrieval", Cambridge University Press (2008).

11. Pavan, Rytaro Ichise, O.P. Vays.:"LiDDM: A Data Mining System for Linked Data":WWW 2011 workshop: Linked Data on the web (LDOW)(2011).

12. Jennifer Neville, Foster Provost." Predictive modeling using social networks":14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008).

13. Jiawei Han, Yizhou Sun, Xifeng Yan, Philip S. Yu.:"Mining Knowledge from Databases: An information Network Analysis Approach": ACM SIGMOD Conference tutorial, Indianapolis, Indiana (2010)

14. Knowledge Discovery Laboratory Dataset. DBLP (http://www.informatik.uni-trier.de/- Iey/db/index.html/). Retrieved Sep. 13, 2010 from DBLP database (2010).

15. Aydoğdu, Ş. (2020). EDUCATIONAL DATA MINING STUDIES IN TURKEY: A SYSTEMATIC REVIEW . Turkish Online Journal of Distance Education , 21 (3) , 170-185 . DOI: 10.17718/tojde.762046.