A Similarity Score based Link Prediction system using Hierarchical Clustering

Zubair Ahmad Lone, Aaqib Iqbal Wani, Prof. Sanjay Sharma

***Abstract*—Link prediction is a significant factor of any social networking site analysis. Social network sites like Facebook and LinkedIn use link prediction as a feature. In this paper, we propose a technique based on node similarity measure and hierarchical clustering. We first calculate the similarity score using Jaccard's coefficient and then assign the score to all pairs of nodes and apply an Artificial Neural Network to the system. This technique improves the accuracy of Link Prediction and is backed by results that prove the proposed model outperforms other methods when both precision and recall of the prediction results are considered.**

***Index Terms*—Hierarchical Clustering, Jaccard's Coefficient, Link prediction, Neural Networks, Similarity Score.**

I. INTRODUCTION

The Link prediction problem is a relevant aspect of social networks related to inferring missing links from an observed network. The additional or missing links may not be directly evident but are most likely to exist when a network of interactions is created on based on the data in hand. It can be useful for researchers and organizations in the areas of artificial intelligence and data mining in identifying unknown interactions within the organization. Link prediction is applicable in other fields such as molecular biology, recommender systems, and criminal investigations.

Social Networks are dynamic and sparse which make the prediction of outcome even more complex. An information system is defined by a network consisting of a set of nodes or vertices and edges or links. In Social networks, the nodes are individuals or a single person and the links correspond to the relationship between the individuals. The communication between individuals is termed as friendship or partnership. Social Network Analysis has attracted researchers from fields of marketing, forecasting, etc. There are several types of networks such as Technological Network, Social Network, Information Network and Biological network [1].

Manuscript received June, 2017.
***Zubair Ahmad Lone**, Department of Computer Science and Engineering, SMVD University, Katra, Jammu, India*
***Aaqib Iqbal Wani**, Department of Computer Science and Engineering, SMVD University, Katra Jammu.*
***Prof. Sanjay Sharma**, Department of Computer Science and Engineering, SMVD University, Katra, Jammu, India*

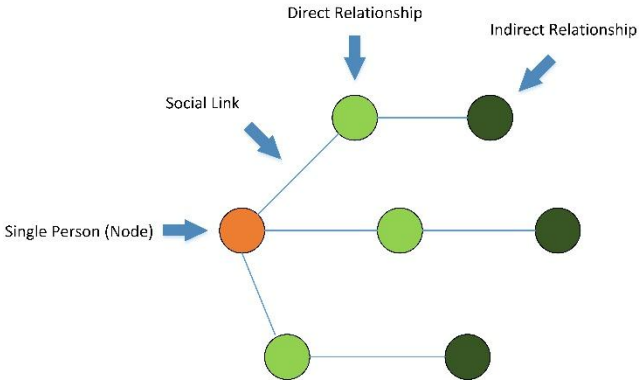

Figure 1 Social Graph Representation

A. Network

A Network can be defined as a set of vertices/nodes interconnected with edges/links. There are various types of networks:

1. Social Network: It is a structure that consists of two components: An individual and the association among them. We can take the example of Facebook where the individual joins the social networks, creates his profile, publishes the content and creates links through friendship, mutual interests, likes, and interests. The meaning of friendship depends upon the type of network [2].
2. Information Network: An information network is a set of at least two or more computers connected to share information and resources like a printer, hard disk etc. it is also called computer network. Computing and telecommunications are the two most important technologies used by information networks. Telecommunication is defined as a process of transferring information over a distance through radio waves, optical signals etc. the network computers are connected with one another through cables, satellite or phone lines [3].
3. Biological Network: A biological network is a network that applies to biological systems. A network is any system with sub-units linked into a whole, such as species characterized as units linked into a whole food web. The mathematical representation of connections as provided by biological networks found in ecological, evolutionary, and physiological studies, such as neural networks. The analysis of those mathematical representations provided by biological networks with respect to human diseases has led to the field of network medicine [4].

B. Link Prediction

Link Prediction is the fundamental problem wherein we predict the future links i.e. links which will occur in the near future. Mathematically it can be defined as "Given a structure

of social network at time t, predict the new links that will be formed in the structure at time t' "[5]. The various strategies for link prediction are:

1. Similarity Based Strategies: Various similarity measures are based on i) Path Distance ii) Common Neighbors iii) Katz Clustering iv) Jaccard Coefficient v) PageRank vi) Adamic/adar for predicting new links in near future.
   Liben-Nowell et al. introduced the baseline for these approaches. The new links would be predicted by ranking all the pairs of the nodes based on their similarity scores. Liu introduced the metric "AuthorRank" that uses normalized weight instead of degree. Lussier proposed the similarity measure termed as "PropFlow" which is based on PageRank. According to this method, we can estimate the similarity between two nodes by evaluating the probability that is a random walk which starts from the node 'x' and ends at node 'y' in 'n' steps or less using link weights. As such there are many methods which are considered as similarity based techniques [6].
2. Maximum Likelihood Algorithm: The Link Prediction problem can be defined as estimating the probability of two nodes that they will have a connection in near future. It needs a technique to estimate the probability. The author predefines some features of network structure. There is need to maximize the likelihood of the network structure then we can calculate the likelihood of non-connected links. The Stochastic block technique uses this approach, according to this technique all nodes are divided into groups and the chance of two nodes to be connected depends on the group to which they belong.
   The Hierarchical structure technique which is based on the hierarchical organization of the network considers the ancestors of nodes for link likelihood prediction. It needs the knowledge of the network structure. Getoor et al., Murata T et al. proposed a method based on graph proximity measures and weights of existing links in a social network. Brandao et al. proposed two new metrics for recommending new collaborations based on social principle (homophily and proximity) [5] [7].
3. Probabilistic Technique/Supervised Learning: The supervised learning technique optimizes the target function to build a technique that comprises of a group of features. We can extract the features from the network. For example, from the link, we can extract features like degree of node, common neighbor etc. (these are the similarity measures). We can also extract the domain related features from co-authorship data, we can have keyword match, affiliation etc. We can then use these feature for link prediction. The chances of future link for non-connected links can be evaluated by the probability i.e. it depends on that how similar a non-connected link is with the existent link. In previous work, Lu, L.et al. discussed the supervised learning approach that produces better results than the similarity based method for Link Prediction. Sharma et al. proposed and implemented Neural Network Approach for Link Prediction in Scholar Social Networks. They consider the co-authorship data set but not the features of the nodes [8].

C. Clustering

Clustering is sometimes considered the most important unsupervised learning problem, the whole concept of clustering deals with finding a structure in a collection of unlabeled data. Clustering can also be defined as "the process of organizing objects into groups whose members are similar in a certain predefined way" [9]. The type of clustering used in this model is Hierarchical Clustering.

Hierarchical clustering is an algorithm that builds a hierarchy of clusters. This algorithm has two approaches: Bottom up and Top down. The bottom-up approach starts with all the data points assigned to a cluster of their own. Then the two nearest clusters are merged into a single cluster. In the end, this algorithm terminates when there is only a single cluster left. In contrary, Top down approach starts with all data points assigned to a single cluster and recursively splitting till each data point is assigned to a cluster of their own [10].

II. PROBLEM STATEMENT

Link prediction is a relatively modern concept, in this era of social networking sites. The efficiency of the networking site is majorly dependent on the connections established day by day and while all this is happening, the social networking sites try to keep the pace by using certain algorithms to make the prediction of the next link of any user in order to provide the user with suitable choices beforehand.

Link prediction suffers from various drawbacks like insufficient information and the inefficiency of various previously proposed algorithms etc. Keeping the various inefficiencies of the algorithms in mind we try to find out a new way by combining two concepts and the information that they provide to help us better gain the information needed and predict the links efficiently. We propose a model that will incorporate the two concepts of similarity score using Jaccard's Coefficient and Hierarchical Clustering.

III. PROPOSED MODEL

We propose a model based on Similarity score between two nodes and Hierarchal clustering. The Similarity score between two nodes is calculated using Jaccard's Index or Jaccard's Coefficient and a cluster is built using distance metric. Then the similarity scores are compared between nodes within the cluster. We simulate our concept by training a neural network over a small sized online dataset. We then calculate the precision and recall and compare them with the base model [11].

First, data is loaded and the vertices and edges are read and saved in the database. We then organize a graph from the data of the loaded nodes and find the relationship connection between the nodes. Distance based hierarchical clustering is performed to find the node-similarity based on degree. Nodes with a high degree of connection within a cluster have a high probability to be connected.

We use Jaccard's coefficient to find the similarity between two nodes:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B are two nodes whose similarity measure is to be calculated.

|A| = degree of A

|B| = degree of B

|A∩B| = common nodes between A and B.

|A∪B| = sum of degrees of two nodes i.e. A and B minus the number of common nodes between A and B.

We will then organize the data in the training set and target set and train the artificial neural network. There are many metrics which define the efficiency of a link prediction algorithm, while some use accuracy and standard deviation, most use precision and recall. In our model, we evaluate the performance in terms of precision and recall [11].

1. Precision: It is defined as the fraction of relevant instances among the retrieved instances. It is sometimes called as positive predictive value. Suppose a computer program for predicting links in a network predicts l number of links correctly and the total number of links generated is equal to N [12]. Then,

   Precision =1/N,

   Where N is total number of links generated and l is the number of links predicted correctly.
2. Recall: The recall is defined as the total number of relevant instances retrieved among the total number of relevant instances in the set. Suppose a computer program for predicting links in a network predicts l number of links correctly and the number of links generated that should be predicted is equal to M [12]. Then,

   Recall = l/M,

   Where M is the number of all the links that should be predicted and l is the number of links predicted correctly.

IV. Simulation and Results

The proposed system is implemented using MATLAB R2016a. In our implementation, a small online data set is used to implement the proposed model. We calculate the precision and recall using the defined formulae and model the results as shown below. It is observed that our proposed model performs better in case of small data sets.

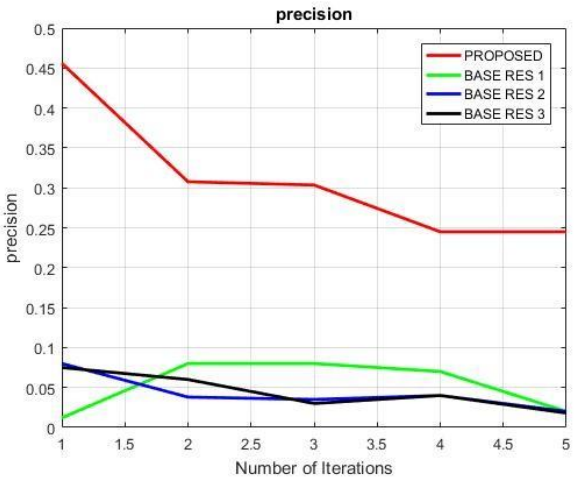

Figure 2 Precision Comparison

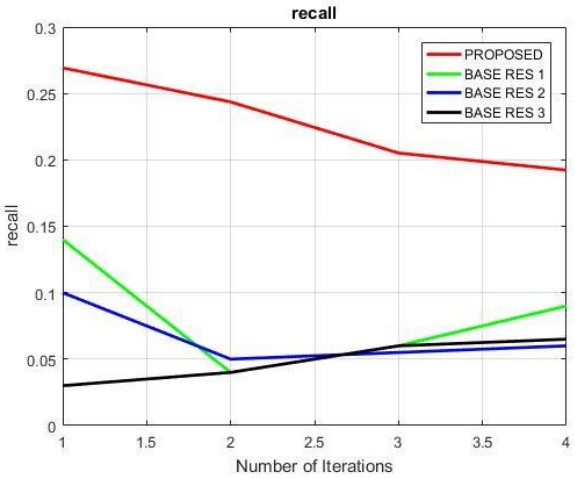

Figure 3 Recall Comparison

V. Conclusion

The proposed model provides better results even when established on relations. Such outcomes establish the ability to use hierarchal clustering and ANN in link prediction. Even though the basic drawback of using hierarchical clustering is that hierarchical clustering becomes messy in the case of large datasets.

A potential future research path will be to analyze social network types and other clustering algorithms and exploit them in providing the information of the network and use that information for better link prediction. Instances of other types of clustering information are content-based like gender, age, and sex to examine our procedures in relation to different fields such as bioinformatics systems.

To avoid one of the most basic disadvantage of hierarchical clustering i.e., implementation on larger data sets, the best thing is to normalize the input and make hierarchical clustering accessible to the link prediction algorithms with larger datasets. Another area to look forward might be using Functional Link Artificial Neural Network.

REFERENCES

[1] Cukierski, William, Benjamin Hamner, and Bo Yang. "Graph-based features for supervised link prediction." Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011.

[2] Smith, M., C. Giraud-Carrier, and Nathan Purser. "Implicit affinity networks and social capital." *Information Technology and Management* 10.2-3 (2009): 123-134.

[3] Hanson, M. David. "The Client/Server Architecture." *Server Management* (2000): 3.

[4] Krause, Jens, D. P. Croft, and Richard James. "Social network theory in the behavioural sciences: potential applications." *Behavioral Ecology and Sociobiology* 62.1 (2007): 15-27.

[5] Krackhardt, David. "Super Strong and Sticky." *Power Influ. Organ* (1998): 21.

[6] Liben- Nowell, David, and Jon Kleinberg. "The link- prediction problem for social networks." *journal of the Association for Information Science and Technology* 58.7 (2007): 1019-1031.

[7] Sharma, Upasana, and Bhawna Minocha. "Link Prediction in Social Networks: A Similarity score based Neural Network Approach." *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. ACM, 2016.

[8] Backstrom, Lars, and Jure Leskovec. "Supervised random walks: predicting and recommending links in social networks." *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011.

[9] Berkhin, Pavel. "A survey of clustering data mining techniques." *Grouping multidimensional data*. Springer Berlin Heidelberg, 2006. 25-71.

[10] Berkhin, Pavel. "A survey of clustering data mining techniques." *Grouping multidimensional data*. Springer Berlin Heidelberg, 2006. 25-71.
[11] Musial-Gabrys, Katarzyna. "Hybrid structure-based link prediction model." Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. IEEE, 2016.
[12] Ting, Kai Ming. "Precision and recall." *Encyclopedia of machine learning*. Springer US, 2011. 781-781.

Zubair Ahmad Lonehas aMasters degree in Computer Science and Engineering from SMVD University, Katra Jammu and a Bachelors degree in Computer Science and Engineering from MIET, Jammu. He has been involved in a number of projects relating to the field and his areas of expertise are Neural Networks and Artificial Intelligence.

Aaqib Iqbal Wanihas aMasters degree in Computer Science and Engineering from SMVD University, Katra Jammu and a Bachelors degree in Computer Science and Engineering from Bharath University, Chennai. His areas of expertise include networking and Cloud computing.

Prof. Sanjay Sharmaholds a Masters degree in Computer Science and Engineering and is currently working as an Assistant Professor in the Department of Computer Science and Engineering at SMVD University, Katra, Jammu. He has been involved in a number of projects relating to the field of Neural Networks and Artificial Intelligence. His areas of expertise include Image Processing, Computer Architecture and Artificial Intelligence.