International Journal of Computer Applications (0975 - 8887)
Volume 121 - No. 2, July 2015

Empirical Analysis of Traditional Link Prediction Methods

Jagadishwari.V
Research Scholar, VTU
Department of Computer Science and Engineering
BMS College of Engineering, Bengaluru

Umadevi. V
Department of Computer Science and Engineering
BMS College of Engineering
Bengaluru

ABSTRACT

Online Social Networks are growing exponentially due to which a lot of researchers are working on Social Network analysis. Link Prediction is a task of predicting new links that may occur in future in the social network. The link prediction problem has generated a lot of interest due its widespread applicability across many domains. We conducted a study on the different methods that have been developed for link prediction. In most of these methods, the social network is modeled as a graph, and the links are predicted based on the similarities between two nodes. We have chosen seven widely used similarity methods in our study. We found that on the simulated data sets, Sorenson index method and Jaccard coefficient method performed well when compared to other methods.

General Terms:

Social Network analysis, Machine learning

Keywords

Link Prediction , Similarity measure

1. INTRODUCTION

A social network is a group of people, organizations or social entities who share something in common, the common bond could be a set of social relationships such as friendship, co-working or information exchange. With the advent of internet, social networking has predominantly become online. People interact on the web using Online Social Networks, such as Facebook and Twitter sharing experiences and knowledge. A social network can be represented as a graph, where nodes represent actors and edges represent ties or interactions. For example collaboration between scientists can be studied using Co-authorship networks [1], with the nodes representing scientists and edges showing the collaboration between them. Co-authorship Networks are academic social networks, Fig 1 shows an example of a Co-authorship network with the nodes representing authors, and the edges representing Co-authorship. The nodes A, B, C, D, and E represent authors and the edges between them represent Co-authorship, i,e, there is an edge between two nodes if the authors have coauthored a paper together. In the Fig 1, A has coauthored with B and C, B has coauthored with A, D has coauthored with C and E and so on. Analyzing social networks reveals some important information that can be used in many applications. Social network Analysis is a field of research that deals with the techniques and strategies for the study of social networks. Link Prediction is one such task. It is a problem of predicting future ties between two entities in a network based on the current relationships. There is a lot of ongoing research on this problem because it is interdisciplinary and has attracted researchers from the fields of physics, economics and computer science [2]. Link Prediction is useful in the study of network evolution [3], it also finds its application in recommendation system on social networks [4], improving sales in E-commerce and finding experts and collaborations in academic Social network [5]. It can be used in biological networks such as protein to protein interaction and metabolic networks. It also finds its applicability in unraveling the unknown connections in the terrorist networks [6][7]. This wide spectrum of application of Link Prediction motivated us to work on this problem. In the literature we find that many methods have been proposed to solve this longstanding problem. We have considered seven methods of Link Prediction in our study, and presented an analysis on how they performed on simulated data sets. The traditional methods will be discussed in detail in section 2. Experimental set up adopted for performance analysis of link prediction methods will be discussed in section 3. In section 4 results obtained will be discussed.

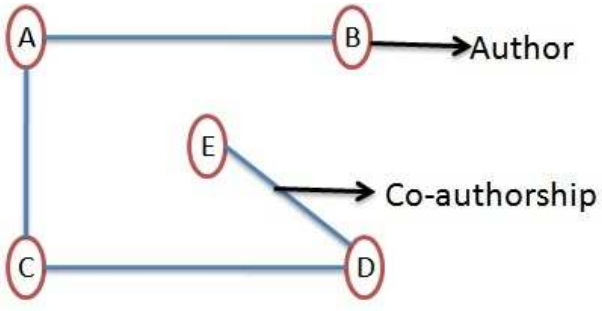

Fig. 1. A Sample Co-authorship Network

2. RELATED WORK

We find a surge in the research on large networks recently. Computational analysis of social network has gained prominence, and link prediction is one such problem. The problem of link prediction was formally stated by libnen-Nowell [8] as ”Given a snapshot of a so-

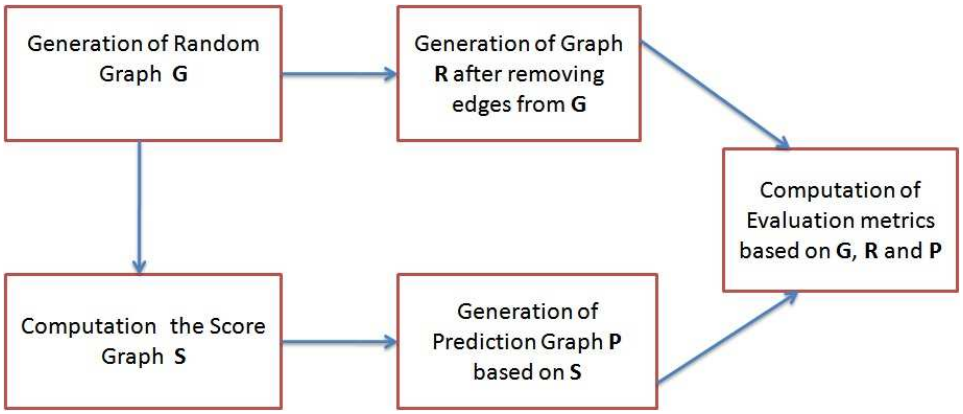

Fig. 2. Experimental Setup for performance analysis study of link prediction methods

cial network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'". Many methods have been proposed in literature for prediction of Links in Social Networks. These networks are modeled as a graph and the traditional methods of Link Prediction are based on the similarity between the two nodes. There are various methods to compute similarity, among them we have considered seven methods in our study. These methods were chosen because they have been extensively studied by most of researchers in the area, and these methods are also used as a baseline to propose other novel methods for Link Prediction.
Link Prediction methods are broadly classified into two categories Methods based on Node Similarities and Methods based on Ensemble of Paths

2.1 Methods based on Node Similarities

For convention in this paper lowercase alphabetical letters are used to represent nodes in the social network and uppercase alphabetical characters are used to represent adjacency matrices of the network. If matrix A is the adjacency matrix of a given social network, then $\Gamma(x)$ denote the set of neighbors of node *x*, and $|\Gamma(x)|$ denote the number of neighbors of node *x*. we discuss six different methods based on Node Similarity

(a) Common Neighbor (CN): The CN [9] is defined as the number of nodes that both *x* and *y* have a direct interaction with. If there are more number of common neighbors between nodes *x* and *y* it makes it more likely that a link can come up between them. This measure is as defined below:

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)|$$

where $\Gamma(x) \cap \Gamma(y)$ represent the intersection of set of neighbors of nodes *x* and *y*

(b) Jaccard Coefficient (JC): It assumes higher values for pairs of nodes which share a higher proportion of common neighbors relative to total number of neighbors they have [10]. This measure is defined as:

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

where $\Gamma(x) \cup \Gamma(y)$ represent the union of set of neighbors of nodes *x* and *y*

(c) Sorensen Index (SI): Besides considering the size of the common neighbors , it also points out that lower degrees of nodes would have higher link likelihood [11]. The measure is defined as:

$$SI = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|}$$

where $|\Gamma(x)| + |\Gamma(y)|$ represent the sum of the number of neighbors of nodes *x* and *y*

(d) Salton Cosine Similarity (SC): SC [12] is a common cosine metric for measuring the similarity between two nodes *x* and *y*. It is defined as:

$$SC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \,.\, |\Gamma(y)|}}$$

where $|\Gamma(x)| \,.\, |\Gamma(y)|$ represent the product of the number of neighbors of nodes *x* and *y*

(e) Adamic-Adar (AA): It was proposed by Adamic and Adar for computing similarity between two web pages [13], subsequently it has been used in Social Networks. It is defined as:

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

(f) Preferential Attachment (PA): The PA [14] metric indicates that new links will be more likely to connect higher-degree nodes than lower ones. It is defined as:

$$PA(x,y) = |\Gamma(x)| \,.\, |\Gamma(y)|$$

2.2 Methods based on Ensemble of Paths

(1) Katz (KA): Katz metric [15] is based on the ensemble of all paths, and it counts all paths of different lengths between two. The shorter paths are given more weights by using damping factor β. This measure is defined as :

$$Katz = \sum_{l=1}^{\infty} \beta^l \left| paths_{x,y}^{l} \right|$$

The similarity measure between every pair of node in the considered network is computed using CN, JC, SI, SC, AA, PA and KA. Based on the similarity measured new links (or future links) will be predicted in the network. The details of this empirical analysis is discussed in the following Section.

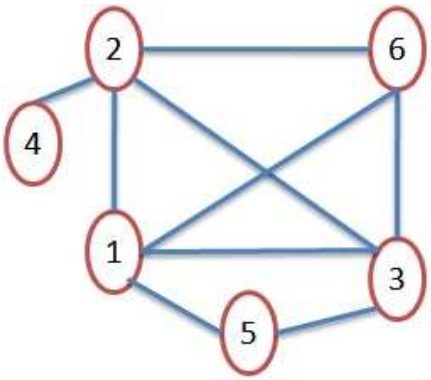

Fig. 3. A sample random Graph **G**

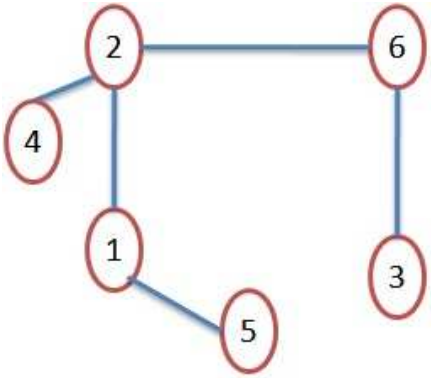

Fig. 4. Graph **R**: After edge removal in **G**

Table 1. CN score for Graph G.

Node Pair	Score(CN)
(1,2)	2
(1,3)	3
(1,4)	1
(1,5)	1
(1,6)	2
(2,3)	2
(2,4)	0
(2,5)	2
(2,6)	2
(3,4)	1
(3,5)	1
(3,6)	2
(4,5)	0
(4,6)	1
(5,6)	2

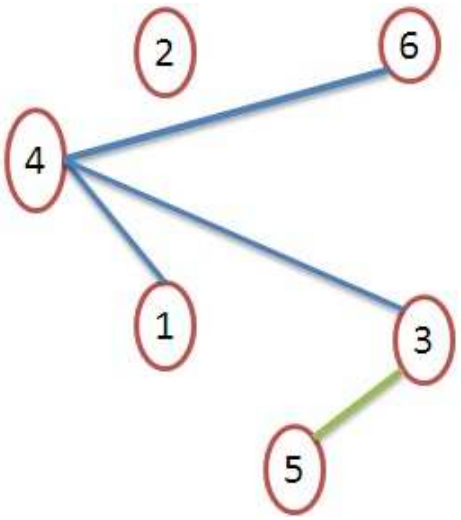

Fig. 5. Graph **P**: Predicted graph for Threshold of one for CN

3. EXPERIMENTAL SET UP AND DATA

The traditional link prediction methods discussed in section 2 was implemented. The experimental set up of this implementation procedure is shown in Fig 2. A graph 'G' is randomly generated, from this graph some of the existing links are removed to generate Graph **'R'**. The score for every node pair in graph **G** is computed based on the various methods discussed in section 2. New links are predicted based on the score to produce the Prediction Graph **'P'**. The evaluation metrics are computed and performance of the methods are analyzed. This process is explained with an graph shown in Fig 3.

The Fig 3, shows a randomly generated graph **G** with 6 nodes and 9 links. Fig 4, shows the graph **R** after removing 4 of the existing links from the initial random graph **G**. The scores are computed for every node pair in the Graph **G** using the Common neighbor method. The score for a node pair (x , y) is the number of common links they share, for example the node pair (1,3) has three common links (i,e both nodes 1 and 3 have a link to nodes 2,5 and 6) which is the score for this node pair. The scores of every node pair is listed in Table 1. All the edges are undirected, hence the score for node pair (x , y) is the same for node pair(y , x). New links are predicted based on the score for each node pair to produce the graph **P**. To predict the new links the threshold is set to the minimum score which is greater than 0, in the example minimum score is 1, so we predict that a link may exists between two nodes that has one common neighbor. The Prediction graph **P** after predicting the links with threshold set to 1, is shown in Fig 5. This graph has four links, one of which was removed in R is correctly predicted, it also predicts three new links that did not exist. It does not predict a link for the node pair(1,5) though the score for this node pair is 1, because a link for this node pair existed in **G** and was not removed in R. The threshold is incremented to the next higher score, links are predicted between the node pairs that has common neighbors equal to this threshold. This process is repeated for threshold ranging from minimum score to the maximum score i,e the number of common neighbors between the considered node pair, in this example the threshold will take three values 1,2, and 3. This process which is described for com-

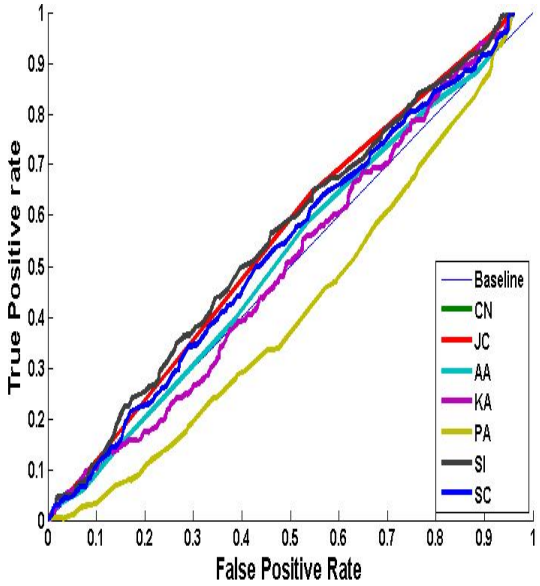

Fig. 6. ROC curve for graph of 50 nodes

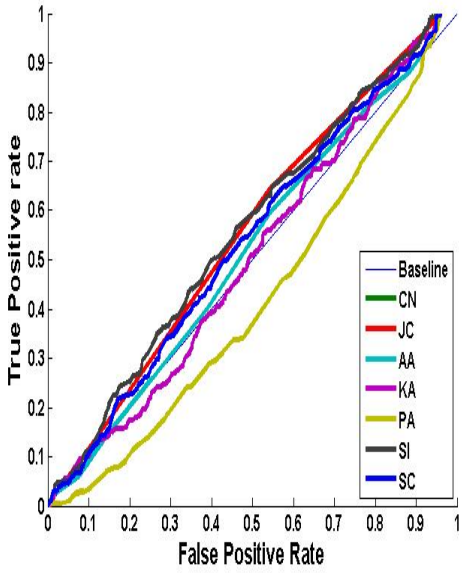

Fig. 7. ROC curve for graph of 100 nodes

mon neighbor method is done for all the other seven methods described in section 2. After predicting the links based on various Link Prediction methods, they were compared by looking at how many of the removed links were correctly predicted. This is done using the Evaluation metrics from machine learning [16]. The details of evaluation metrics is discussed in the next section.

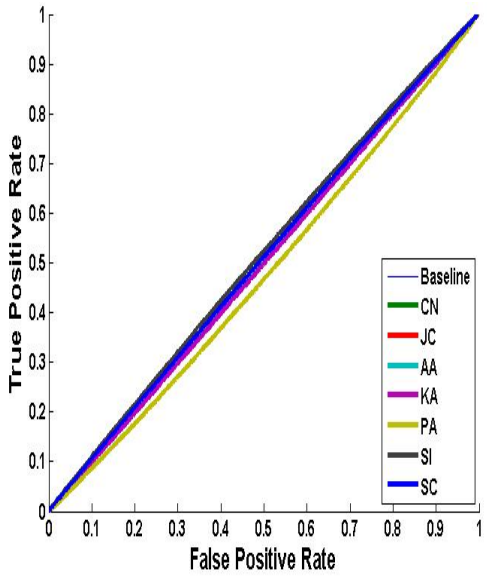

Fig. 8. ROC curve for graph of 1000 nodes

4. RESULTS AND DISCUSSION

We will introduce the evaluation metrics that was used to gauge the performance of the seven link prediction methods. Terminologies of the evaluation metrics are as follows:

Positive: For a node pair (x, y) if there exists already a link between them, such a link is referred to as positive.

Negative: For a node pair(x,y), if there is no link between them we say there is a negative link.

True Positive (TP): When the removed link is been predicted by a Link Prediction method then it is referred as a "True Positive Link"

False Positive (FP): When the link is been predicted by a Link Prediction method, but the link was not removed then it is referred as a "False Positive Link"

True Negatives (TN): When the link was not removed, and has not been predicted by a Link Prediction method then it is referred as a "True Negative Link"

False Negatives(FN): When the removed link was not been predicted by a Link Prediction method then it is referred as a "False Negative Link"

The True positive rate (TPR) and the False positives rate (FPR) is computed as follows:

Let

N_{TP}:Represent Number of true positives

N_{FP}:Represent Number of false positives

N_{TN}:Represent Number of true negatives

N_{FN}:Represent Number of false negatives

Then,

$$TPR = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

$$FPR = \frac{N_{FP}}{N_{FP} + N_{TN}}$$

ROC (Receiver Operating Characteristic): The ROC curve plots TPR against FPR. A prediction metric is said to be better

than any other prediction metric if the ROC curve is consistently higher. Fig 6 shows the ROC for a random graph of 50 nodes, Jaccard Coefficient (JC) method performs better than other methods, Fig 7 shows the ROC for 100 nodes, Sorenson Index (SI) gives a better performance. Fig 8 shows the ROC of 1000 nodes, we can still find Sorenson Index performed slightly better than other methods. Preferential Attachment (PA) has a below average performance in all three cases.

5. CONCLUSION

In this paper, seven similarity methods that is widely used for link prediction was chosen for study. All these methods were implemented and their behavior on simulated data was analyzed. On the simulated data sets, Jaccard Coefficient method performed better when compared to other methods for a network of 50 nodes . The Sorenson Index performs slightly better than Jaccard Coefficient when the number of nodes are 100 and 1000. This study has given us an in-depth understanding of the problem of link prediction. We have identified the gaps in the existing Link Prediction methods. Most of the methods take a static view of the network, but social networks are highly dynamic. Social networks are growing exponentially and hence scalability is another issue in most of the existing methods. In future we intend to address these issues and propose novel method for link prediction with improved performance.

6. ACKNOWLEDGMENT

The work reported in this paper is supported by the college through the Technical Education Quality Improvement Programme (TEQIP-II) of the MHRD, Government of India.

7. REFERENCES

[1] Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205, 2004.

[2] Peng Wang, BaoWen Xu, YuRong Wu, and Xiaoyu Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, pages 1–38, 2014.

[3] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[4] Tsung-Ting Kuo, Rui Yan, Yu-Yang Huang, Perng-Hwa Kung, and Shou-De Lin. Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 775–783. ACM, 2013.

[5] Milen Pavlov and Ryutaro Ichise. Finding experts by link prediction in co-authorship networks. *Finding Experts on the Web with Semantics*, 290:42–55, 2007.

[6] Steve Ressler. Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs*, 2(2):1–10, 2006.

[7] Stuart Koschade. A social network analysis of jemaah islamiyah: The applications to counterterrorism and intelligence. *Studies in Conflict & Terrorism*, 29(6):559–575, 2006.

[8] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[9] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[10] Paul Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[11] Thorvald Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34, 1948.

[12] Salton, Gerard, McGill, and Michael J. Introduction to modern information retrieval. 1983.

[13] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. 286(5439):509–512, 1999.

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

[16] Zhifeng Bao, Yong Zeng, and YC Tay. Sonlp: Social network link prediction by principal component regression. In *Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference on*, pages 364–371. IEEE, 2013.