An Experimental Comparison of the Link Prediction Techniques in Social Networks

D. Sharma, U. Sharma, and Sunil Kumar Khatri

***Abstract*—Link prediction is a very well studied problem as it has applications in many different areas. Many algorithms have been presented in the literature for the Link prediction problem. The general for of the problem is that given the topology of graph G at a certain time t, we need to predict the topology of the graph *G* at time *t*' where *t*' > *t* assuming that the number of nodes does not change. The techniques used for Link prediction are categorized into three types: Nodes based techniques, Link based techniques and Path based techniques. Then there are other techniques that use meta-approaches such as.....which are based on the basic techniques. In this paper we conduct a survey of all the existing Link Prediction techniques to the best of our knowledge and perform an experimental comparison of these techniques. We use real social network data for the testing.**

***Index Terms*—Link prediction, social networks.**

I. Introduction

A Social network consists of a group of people and connections between them. These connections can be any type of social link that makes a relationship between two people. This relationship can be represented by nodes and links. Social networks are popular way to model the interactions among the people in a group or community. Social networks are highly dynamic in nature. They grow and change as time changes. They can be visualized as graphs, where a vertex corresponds to a person in some group and an edge represents some form of association between the corresponding persons. The associations are usually driven by mutual interests that are intrinsic to a group. New nodes may appear in the network and new edges may appear to show new interaction in the network. Link Prediction is a very important problem that is an aspect of Social network analysis. The nodes in a sociogram are linked in a complex web of relationships that change over time. These relationships emerge, strengthen and decay as a result of individual's positions in the network, their behaviour and the influence of the environment. Predicting changes to a social network is called the link prediction problem. Link prediction problem is usually described as a task to predict how likely a link exists between an arbitrary pair of nodes. Link prediction is the problem of identifying whether a link exists between two objects. There are many application areas where Link prediction is applicable. In the area of internet and web science, tasks like automatic web hyperlink creation, website hyper-link prediction. Link prediction is used to build recommendation systems in e-commerce. The other applications are in bibliography, library science and de-duplication. The link problem has been more formally defined as both the identification of unobserved links in a current network or as a time series problem where the task is to predict which links will be present in the network at a time $t + 1$ given the state of a network at time *t*. Link prediction can be described as such two questions. First, given a network at some point, can we get the new interactions among its members which are more likely to happen second, can we infer some missing interlinks which can not be observed in the network, though they have some real connections Liben-Nowell and Kleinberg [1] proposed a model for link prediction based on node similarity. There are several categories of node similarity, one is the neighbourhood based similarity like common neighbors of two nodes, the other one is path based similarity like shortest path distance of two nodes. So Link Prediction can be categorized into two classes: (1) Problem of identifying existing yet unknown links. (2) Predicting links that may appear in the future.

In this paper, we are interested in predicting links that may appear in the future. The Link Prediction problem can be formally aefined as follows: Given a snapshot of the topology of a social network at time t, we need to predict the topology of the graph *G* at time *t*' where $t' > t$ assuming that the number of nodes does not change. It is assumed that the edges do not carry any weights.

There have been numerous techniques proposed for Link Prediction problem. The techniques may be based on Graph theoretic approach, Statistical approach, Supervised learning approach, Clustering. In this paper, we perform an experimental comparison of the different Link Prediction techniques. In this study, we test the performance of the different techniques based on the Precision. The paper is divided into the following sections: Section I contains the Introduction, Section II contains the Background, Section III contains the Experiments, Section IV consists of the Results and Analysis and Section V consists of the Conclusion and future work.

II. Background

Liben-Nowell and Kleinberg [1] proposed one of the earliest link prediction models that works explicitly on a social network. Every vertex in the graph represents a person and an edge between two vertices represents the interaction between the persons. The association are usually driven by mutual interests that are intrinsic to a group. Multiplicity of interactions can be modelled explicitly by allowing parallel edges. The predictors designed for Link

Manuscript received August 12, 2013; revised October 23, 2013.

The authors are with Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, India (e-mail: {dsharma10, usharma}@amity.edu, sunilkkhatri@gmail.com).

Prediction can be broadly classified into three categories, namely, predictors based on graph distance, predictors based on node neighbourhoods and predictors based on path topology. The predictors can be viewed as techniques used to measure the 'proximity' or 'similarity' between two nodes *x* and *y* relative to the network topology. Some common approach in predicting links with the help of graph topology is based on the following common behaviour: It is observed that people who are close in the network have friends in common and travel in similar circles. They are more likely to connect in the near future. The graph theoretic approach is divided into three categories: Techniques based on Node neighbourhood, Path based techniques and Distance based techniques.

A. Graph Distance Predictor

The basic approach for measuring node proximity in social network by measuring the graph distance between them i.e. we find the pairs (*x*, *y*) by the length of the shortest path connecting them in graph. So shortest path predictor selects a random subset of distance-two pairs.

B. Predictors Based on Node Neighborhoods

This method is based on the idea that the two nodes *x* and *y* are likely to form a link in the future if their sets of neighbours have Γx and Γy have large overlap. An degree (*x*, *y*) is more likely to form if edges(*x*, *z*) and (*z*, *y*) are already present for some *z*.

C. Common Neighbors [2]

Is a Node Neighborhood based technique. For two nodes, *x* and *y*, the size of their common neighborhood is defined as $|\Gamma x \cap \Gamma y|$, where Γx is the set of neighbours of *x* and Γy is the set of neighbours of *y*. This technique is based on the intuition that if there is a node that is connected to *x* as well as *y*, then there is high probability that vertex *x* be connected to vertex *y*. Thus, as the number of common neighbours grow higher, the probability that *x* and *y* have link between them increases. In other words two nodes *x* and *y* are more likely to have a link if they have many common neighbors. Newman [1] has computed this quantity in the context of collaboration Networks and used this predictor to compute the possibility that two authors will collaborate in the future in co-authorship networks.

D. Jaccard Coefficient [3]

This was proposed by Jaccard and it presents a normalized form of Common Neighbor technique. It is based on the logic that a judgement cannot be made simply based on the Common Neighbors, but a Normalized value should be taken. Jaccard Coefficient normalizes the size of common neighbors as below:

$$\text{Jaccard-coefficient } (x, y) = \frac{|r(x) \cap r(y)|}{|r(x) \cup r(y)|}$$

It defines the probability that a common neighbour of a pair of vertices *x* and *y* would be selected if the selection is made randomly from the union of the neighbour-sets of *x* and *y*. So for high number of common neighbors, the score would be higher.

E. Adamic/Adar [4]

This technique was firstly proposed for the metric of similarity between two web pages. It calculates the probability when two personal homepages are strongly related. It computes features that are shared among nodes and then defines the similarity between them. In case of Link Prediction in Social networks using only topological information, the features are Neighbours. This predictor depress the power of high-degree common neighbors because that high-degree nodes are usually stars of the network and the nodes connected with these stars may hardly know each other. For this first the features of the pages are computed and then the similarities are defined.

$$\text{Score}(x, y): \sum_{z} \in r(x) \cap r(y) 1/\log|r(z)|$$

So Adamic/Adar weighs the common neighbors with smaller degree more heavily.

F. Preferential Attachment [5]

Is based on the fact that the probability that a new edge is added to the network with node *x* as an endpoint is proportional to $|\Gamma x|$, that is the number of neighbours of *x*. If we consider the neighbourhood size as feature value, then multiplication can be an aggregation function, which is named as preferential attachment score:

$$\text{Score}(x, y) = |\Gamma x|, |\Gamma y|$$

G. Katz [6]

This technique defines a measure that searches for the set of all paths from *x* to *y* node and sums them or it defines the measure that directly sums over collection of the paths, exponentially damped by length to count short path more heavily. Where paths $\overset{<1>}{x}, y$ are the set of all length – l paths from *x* to *y*, and $\hat{a}$ is a pre-defined constant.

$$\text{Score } (x, y): = \sum_{i=1}^{\infty} \beta^{1} \left\| paths \overset{<1>}{x}, y \right|$$

H. Hitting Time [7]

Hitting Time is designed in context of random walks on a graph**.** A random walk on graph starts at a node *x* and iteratively moves to a neighbour of *x* chosen uniformly at random from the set Γx. The hitting time $H_{x, y}$ from *x* to *y* is the expected number of steps required for a random walk starting at *x* to reach *y*. If hitting time is less, it means nodes are similar to each other. So chances are more to have link in future.

I. Commute time [8]

This is calculated as Score (*x*, *y*)= *Hx*, *y* + *Hy*, *x*, Hitting Time metric is not symmetric so for undirected graph, Commute time can be used.

J. Rooted Pagerank [9]

It is used for web-page ranking and has inherent

relationship with the hitting time. So pagerank value can also be used as a feature of link prediction. It measures the Score (x, y) as the stationary probability of y in a random walk that returns to x with probability in each step and probability (1- α) that it will move to a different neighbour. The pagerank is the attribute of single vertex so there is need to modify so that it can show similarity between pair of vertices.

K. LRW and SRW [10]

These are techniques that use the concept of local random walks. Local Random walk is a technique with lower computational complexity compared to Random Walk with Restart or Average Commute Time. Random walk is a Markov chain describing the sequence of nodes visited by a random walker. The process can be described by P (Transition Probability Matrix), where $P_{x,\ y}$ presents the probability that the random walker staying at node x will walk to y in the next step. $P_{x,\ y} = a_{x,\ y}/k_x$, where $a_{x,\ y}$ =0 is x and y are not connected and $a_{x,\ y}$ = 1 if x and y are connected. The term k_x denotes the degree of the node x. LRW is the abbreviation for local random walk and SRW is t for supervised random walk. One problem that may occur in LRW is that x and y may be close to each other, but there is a chance that the random walker may go too far from x and y. In this case the closeness between x and y may be incorrectly estimated. SRW uses the concept of continuously releasing the walkers at the starting point, resulting in a high similarity between target node and the nearby nodes.

L. Simrank [11]

Simrank is based on the following : If two neighbours are so close to each other that they should be joined by an edge, then Similarity Score (x, y) is calculated as:

$$y \frac{\sum_{\alpha \varepsilon \tau(x)} \sum_{b \varepsilon \tau(y)} Similarity(a,b)}{|\tau(a)||\tau(b)|}$$

III. Experiments

We perform a comparison of all the techniques that have been presented in the previous section. To perform the tests, we used the Epinion [12] data. Epinion is a general consumer review website where members post their review for large range of products and belongs to the who-trusts-whom category of online social networks. The network has 75,879 nodes and 508837 edges. The diameter of the graph is 13.

To perform the experiments, we create a test data by removing 10% of the edges from the graph, which is a normal practice in Social network analysis. We test twelve existing Link Prediction Techniques to perform the comparison. The techniques compared are Node Neighbourhood, Jaccard's Co-efficient, Adamic/Adar, Hitting Time, Preferential Attachment, Katz, SimRank, Commute Time, Normalized Commute Time, LRW, SRW and Rooted Pagerank. We vary the value of â for Katz and the value of á for Rooted Pagerank. To measure the performance, we calculate the Precision of each technique. The precision is the ratio of the correct edges identified to the total number of edges identified by the technique. To perform the comparison, we implemented the techniques using C++ programming language and gcc compiler. The platform used was a Intel Core2Duo processor with 4 GB RAM and 2 GHz processor speed.

It is observed from the results in Table I that LRW algorithm shows the best performance. However, Node Neighbourhood, Jaccard's Coefficient and SRW are also close. The performance of Katz and Rooted Pagerank show a huge difference when the values of the â are varied. But overall we can conclude that LRW shows best performance among all the Link prediction techniques.

TABLE I: Performance of Various Link Prediction Techniques

No.	Techniques	Precision
1	Node Neighbourhood	16.4
2	Jaccard's Coefficient	15.8
3	Adamic/Adar	15.1
4	Hitting Time	5.9
5	Preferential Attachment	7.1
6	Katz ($\hat{a}$ = 0.01)	6.4
	Katz ($\hat{a}$ = 0.001)	11.9
	Katz ($\hat{a}$ = 0.0001)	17.8
7	Sim Rank	15
8	Commute Time	4.6
9	Normalized Commute Time	5.1
10	LRW	18.2
11	SRW	16.7
12	Rooted Pagerank ($á$ = .01)	9.7
	Rooted Pagerank ($á$ = .1)	13.9
	Rooted Pagerank ($á$ = .5)	18.3

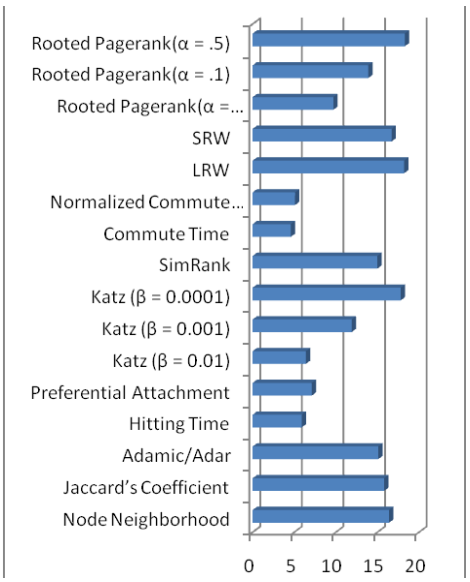

Fig. 1. Graph representing the performance of various link prediction techniques.

IV. Conclusion and Future Work

We compared 12 techniques for Link Prediction on a real dataset. It was observed that many of the techniques

performed well, but LRW shows the best performance. Rooted Pagerank with $\alpha = 0.5$ is close, but the performance of Rooted Pagerank varies a lot with different values of α. Thus, we need a very good estimate of the value of α that needs to be used in the technique to make a good prediction. As we can conclude from the results, there is still a scope of improvement in the performance and new techniques need to be designed to improve the Precision to close to 100%.

ACKNOWLEDGEMENT

The authors thank Dr. Ashok K. Chauhan, Founder President, Amity University, for his support and encouragement along with providing us the necessary infrastructure for research.

REFERENCES

[1] D. L. Nowell and J. Kleinberg, "The link-prediction problem for social network," *Journal of the American Society for information science and Technology*, vol. 58, no. 7, pp. 1019-1031, 2007.
[2] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Physical review letters E*, vol. 64, 2001.
[3] P. Jaccard, *Bulletin De La Societe Vaudoise Des Science Naturelles*, Nabu Press, vol. 37, no. 547, 1901.
[4] L. A. Adamic and E. Adar, "Predicting missing links via local information," *Social Networks*, vol. 25, no. 3, pp. 211-230, July 2003
[5] M. E. J. Newman, "Clustering & preferential attachment in growing networks," *Physical review letters E*, vol. 64, 2001.
[6] L. Katz, "A new status index derived from sociamatric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39 – 43, March 1953.
[7] F. Gobel and A. Jagers, "Random Walks on Graphs," *Stochastic Processes and Their Applications*, vol. 2, no. 4, pp. 311-316, 1974.
[8] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "A link analysis extension of correspondence analysis for mining relational databases," *IEEE Trans. Knowl. Data Eng*, pp. 481-495, 2007.
[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks*, vol. 3, no. 1-7, pp. 107-117, 1998.
[10] W. Liu and L. Lü, "Link prediction based on local random walk," *Europhysics Letters*, no. 5, 2010.
[11] G. Jeh and J. Widom, "SimRank: A Measure of Structural Context Similarity," in *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 538-543.
[12] M. Richardson, R. Agrawal, and P. Domingos, "Trust Management for the Semantic Web," in *Proc. International Semantic Web Conference*, 2003, pp. 351-368.

Dolly Sharma has received Ph.D. in computer science and engineering from University of Connecticut, USA in 2010, and M.Sc. in computer science from Banastyhali University in 2004, India. She got her B.Sc. on computer science from BIT, Mesra, Ranchi, India in 2002. She was an assistant professor at Amity Institute of Information Technology, Amity University. Her research interest is in the areas of advanced algorithms, data mining, computational biology and bioinformatics.

Upasana Sharma has received Ph.D. scholar at Amity University. She is an MCA from UPTU, Lucknow, India and gets B.Sc. (PCM) from Chaudhary Charan Singh University, Meerut, India. She works as an assistant professor at Amity Institute of Information Technology, Amity University. She has 10 years of teaching experience. Her research interest is in data mining.

Sunil Kumar Khatri is working as the director in Amity Institute of Information Technology, Amity University, Noida, India. He is a fellow of IETE, Sr. Member of IACSIT, Sr. life member of Computer Society of India and member of IEEE and IEEE Computer Society. He has been conferred "IT Innovation & Excellence Award for Contribution in the field of IT and Computer Science Education" by *Knowledge Resource Development & Welfare Group* on him during Seminar on Advancement & Outreach of Information Technology: Introspection & the road ahead at IIT, Delhi in Dec. 2012. He has also been conferred with the award for "Exceptional Leadership and Dedication in Research" during the *4th International Conference on Quality, Reliability and Infocom Technology in the year 2009*. He is associate editor of International Journal of Systems Assurance Engineering and Management (IJSAEM), Springer Verlag. He is in Editorial Board of International Journal of Computer Theory and Engineering (Singapore) and International Journal of Modeling and Optimization (Singapore). He has edited three books, two special issues and published several papers in international and national journals and proceedings. His areas of research are software reliability, modeling and optimization, data mining and warehousing, network security, soft computing and pattern recognition.