A Survey On Link Prediction Problem In Social Network

Pallavi Gupta[1], Sanjiv Sharma[2]

[1]M.Tech. (Computer Science & Engineering) Research Scholar M.I.T.S. Gwalior, MP, India

[2] Asst. Prof., Department of CSE & IT M.I.T.S. Gwalior, MP, India

Abstract- *Social network analysis is an emerging field of research and link prediction problem plays an important role for prediction of social network structure. This research paper focuses on existing researches in link prediction problem. Existing researches reveals that link prediction problem complexity, available solutions efficient group communication management and social link awareness. The link prediction problem across aligned networks can include anchor link prediction problem and link transfer across aligned heterogeneous networks. This paper summaries recent progress about link prediction algorithms and survey of all the existing link prediction techniques.*

Index Terms- *Data mining, Link prediction, Social network analysis, Social network.*

I. INTRODUCTION

Social networks are graph structures whose nodes or vertices describe people or other entities grouped in a social context, and whose edges describe interaction or collaboration between these entities. 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. The shape of a social network helps to determine a network's usefulness to its individuals.

This paper defines and studies a basic computational problem underlying social network evolution. Link prediction is a very important problem that is an aspect of social network analysis. Predicting changes to a social network is known as link prediction problem. The link prediction 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. As an example, consider a social network of co-authorship among scientists who are 'close' in the network may be likely to collaborate in the future. Hence, link prediction can be thought of as a contribution to the study of social network evolution models. Link prediction is the only sub-field of social network analysis which has focus on edges between objects. Due to this reason, link prediction become interesting than the traditional data mining areas which focus on objects. Link prediction can be used in various areas like recommender systems and criminal investigations. Approaches to link prediction have been proposed based on several measures for analyzing the "proximity" of nodes in a network lead to the most accurate link predictions. Many measures originate from techniques in graph theory and social network analysis. In link prediction methodology all methods assign a connection weight score (x, y) to pairs of nodes x and y, based on given proximity measure and input graph G. A ranked list in decreasing order of score (x, y) is produced. This gives the predicted new links in decreasing order of confidence. The prediction can be evaluated based on real observations on experimental data sets.

II. BACKGROUND AND RELATED WORK

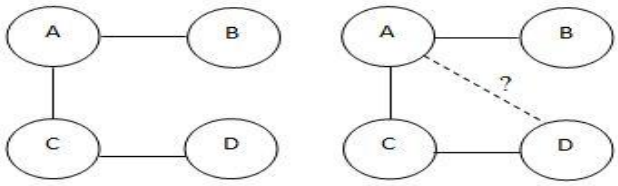
Data mining [1] refers to extracting or "mining" knowledge from large amounts of data. The term data mining should have been more appropriately named as "Knowledge mining from data" or "Knowledge mining". Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. To extract the useful information or knowledge from the stored data is the overall goal of data mining. In Data mining [2] there is analysis of large quantities of data in order to discover meaningful patterns and rules. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns. Data Mining [3] is about solving problems by analyzing data already present in databases. The functionalities of data mining are used to specify the types of patterns to be found in data mining tasks. Data mining tasks can be classified into two categories-descriptive and predictive. Descriptive mining tasks [4] focus on general properties of the data in database. Predictive mining tasks focus on the current data in order to make predictions. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model. The descriptive model works for the main characteristics of the data set and it follows the bottom-up

approach which is the characteristic of undirected data mining. By bottom-up approach means that data itself determines the relationship whether it is useful or not. Predictive model is to predict the target variable when the target variable belongs to predefined class or labels, then it is called classification but it called regression when target variable is real number classification.

Graphs [5] become increasingly important in modeling complex structures like chemical compounds, circuits, images and social networks. The graph representation is used in pattern recognition and machine learning. Many graph search algorithms have been developed in video indexing, computer vision and text retrieval. Graph mining has become an important technique because of the increasing demand on the analysis of large amounts of structured data in data mining. There are various kinds of graph patterns, frequent sub structures are the very basic patterns that can be discovered in a collection of graphs. They are useful for discriminating different groups of graphs, classifying and clustering graphs, characterizing graph sets, facilitating similarity search and building graph indices in graph databases. Many graph mining methods have been developed till now discover the interesting patterns in various applications. Frequent pattern mining is an important part of graph mining that helps to discover patterns that conceptually represent relations among different entities. To develop algorithms for frequently occurring sub-graph from large data set graph is challenging task and complex to compute, as graph and sub graph isomorphism play a key role throughout the computations. The frequent patterns in graph mining can be determined apriori based approach and pattern growth approach. Apriori based approach uses generate and test approach in which it generates candidate item sets and tests if they are frequent. In this one is generation of candidate item sets is expensive (in both space and time) and second support counting is expensive that is subset checking, multiple database scans (I/O) whereas pattern growth approach discovers frequent item set without generating candidate which has two steps, first is to build a compact data structure called the frequent pattern tree and second is extracts frequent item sets from the frequent pattern tree.

A Social network [6] 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. 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 and they can be visualized as graphs, in which a vertex denoted as person in some group and link represents some form of association between the corresponding persons [7]. 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.

Define and study a basic computational problem underlying social network evolution is the link prediction problem:



At Time t At Time t′

Fig 1: Social network graphs G at time t and t′

Given a figure of the topology of a social network at time t, this paper search to accurately predict the edges that will be added to the network during the interval from t to a given future time t' need to predict the topology of thegraph G at timet' where t'>t assuming that the number of nodes does not change.

There have been numerous techniques proposed for link prediction problem. The techniques may be based on supervised learning approach, clustering, graph theoretic approach, and statistical approach. Link prediction problem is usually described as a task to predict how likely a link exists between an arbitrary pair of nodes. In other words 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 web science and internet, tasks like website hyper-link prediction and automatic web hyperlink creation. The use of link prediction in e-commerce is to build recommendation systems.

Lada. A. Adamic and Eytan Adar [8] proposed the metric of similarity between two 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 in which 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. Liben-Nowell and Kleinberg [9] introduced a model for link prediction based on node similarity. There are several categories of node similarity. First one is the neighbourhood based similarity like common neighbours of two nodes and the other one similarity based

on path which tries to determine shortest path distance between two nodes. So link prediction can be categorized into two classes, first is to problem of identifying existing yet unknown links and predicting links that may appear in the future. M. E. J. Newman [10-12] used the concept of clustering & preferential attachment in growing networks. 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. Glen Jeh and Jennifer Widom [13] proposed the concept of Simrank. If two neighbours are so close to each other that they should be joined by an edge. Simrank define a new similarity measure to determine a way of expressing the proximity among graph nodes. Let vi and vj be two graph nodes and sim (vi,vj) a function that expresses their similarity. The probability of two persons being friends becomes higher when the similarity score between two nodes become higher. Liu and Lu [14] in 2010 introduced a link prediction model based on the similarity of the nodes. This is important in applications that consider the similarity of nodes such as gender age etc. For this they introduced two similarity indices based on random walk.

III. AVAILABLE FRAMEWORK FOR LINK PREDICTION

The four different problems [15] given by link prediction are shown in figure below.

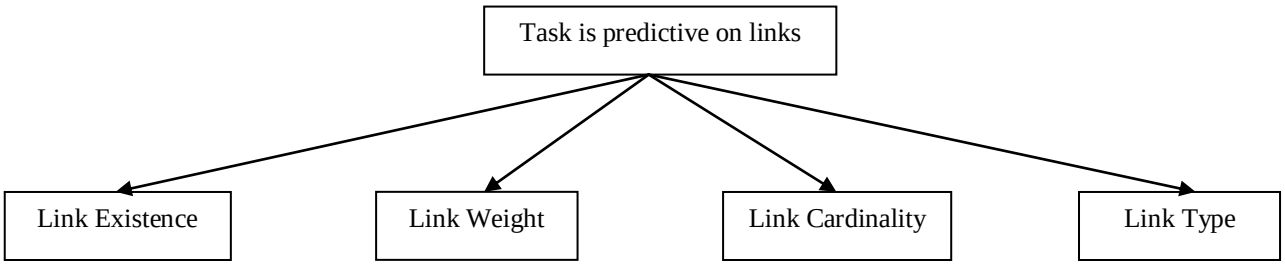


Fig 2: Differentiation of link prediction tasks

Most of the research papers on link prediction focus on problem of link existence (whether a [new] link between two nodes in a social network will exist in the future or not)**.** This is because the link existence problem can be easily extended to the other two problems of link weight (links have different weights associated with them) and link cardinality (more than one link between same pair of nodes in a social network) and the last problem of link type prediction is a bit different which gives different roles to relationship between two objects. This paper contains a framework in which the problem of link prediction id tackled as a binary classification problem. Classification whether a link exists or not can be performed using various classification algorithms like decision tree and support vector machine (SVM).Different features like topological features, content/semantic information of individual nodes can be used for analyzing the proximity of nodes in a social network. This feature information combined with different approaches such as relational data approach and clustering approach help us to predict the existence of a link between two nodes.

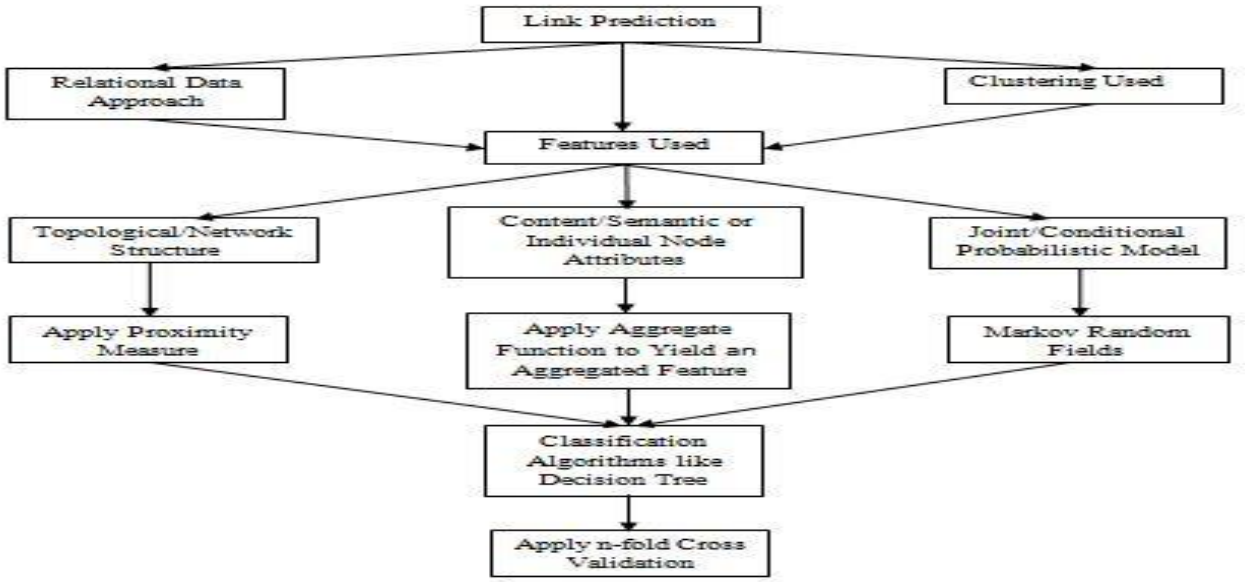


Fig 3: Classification of link prediction approaches

IV. COMPARISON OF THE BASIC LINK PREDICTION METHODS

In this figure, for each vertex $v \in V'$.The V'^2represents the possible edges in$G$".E$'$ contains both E" and the deleted edges. E_{new} represents the k top scoring edges resulting from running a link prediction heuristic. The intersection of $E' - E$" and E_{new} represents the set of successful edge predictions and wish to maximize the size of this set. The upper size limit for this success area is$|E' - E"|$.

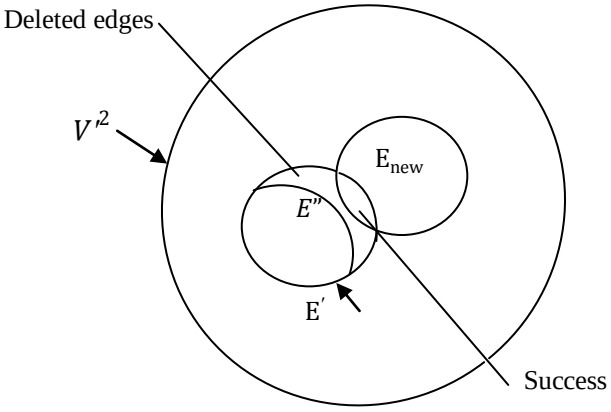


Fig 4: Heuristic link prediction methods

Methods	Function	Basic Implementation	Running Time
Shortest path	$-dx,y$	BFS	$O(V'.n^l)$
Common Neighbours	$\|\Gamma x \cap \Gamma y\|$	List Comparison	$O(V'^2.nlogn)$
Katz	$\sum_{l=1}^{\infty} \beta^l.\|paths_{xy}^{<l>}\|$	DFS	$O(V'.n^l)$
Simrank	$\gamma.\frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} similarity(a,b)}{\|\Gamma(x)\|.\|\Gamma(y)\|}$	Fixed Point Iteration	$O(KV'^2.n^2)$

Table 1: Comparison of the basic link prediction methods with respect to function, approaches and running time

V. MEASUREMENTS OF LINKS FOR PREDICTION IN GRAPH MINING

The framework of link prediction algorithm is based on the similarity of the algorithm where each pair of nodes x and y, is assigned a score S_{xy} .This function is defined as the similarity between nodes x and y. Here this paper introduces some simple link prediction similarity indices.

A. Local Similarity Indices

a. Common Neighbours

Common neighbour is a node neighbourhood based technique. The size of common neighbourhood of two nodes x and y can be defined as

$$S_{xy}^{CN} = |\Gamma x \cap \Gamma y| \dots\dots\dots (1)$$

Equation (1) represent the number of neighbours that x and y have in common. 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 has computed this quantity in the context of collaboration networks, verifying a co-relation between the number of common neighbours of x and y at time t and used this predictor to compute the possibility that two authors will collaborate in the future in co-authorship networks. Kossinets and Watts [16] works to analyze a large-scale social network like face book. In their work, they suggest that two individuals having many common friends are very probable to be friend in the future.

b. Salton Index

Salton index [17] is defined as

$$S_{xy}^{Salton} = \frac{|\Gamma x \cap \Gamma y|}{\sqrt{k_x * k_y}} \dots\dots\dots (2)$$

Wherek_x is the degree of node x and k_y represent the degree of node y. This index is also called the cosine similarity.

c. Sorensen Index

Sorensen Index is defined as

$$S_{xy}^{Sorensen} = \frac{2|\Gamma x \cap \Gamma y|}{k_x + k_y} \dots\dots\dots (3)$$

This index [18] is used mainly for ecological community data.

d. Hub Promoted Index (HPI)

This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks, and is defined as

$$S_{xy}^{HPI} = \frac{|\Gamma x \cap \Gamma y|}{\min\{k_x, k_y\}} \dots\dots\dots (4)$$

In this measurement, the links adjacent to hubs are likely to be assigned high scores since the denominator is decided by the lower degree only.

e. Hub Depressed Index (HDI)

Analogously to the HPI, the HDI also consider a measurement with the opposite effect on hubs. It is defined as

$$S_{xy}^{HDI} = \frac{|\Gamma x \cap \Gamma y|}{\max\{k_x, k_y\}} \dots\dots\dots (5)$$

f. Leicht-Holme-Newman Index (LHN1)

This index assigns high similarity to node pairs that have many common neighbors compared not to the possible maximum, but to the expected number of such neighbors**.** It is defined as

$$S_{xy}^{LHNI} = \frac{|\Gamma x \cap \Gamma y|}{k_x \times k_y} \dots\dots\dots (6)$$

Where the denominator $k_x \times k_y$ is proportional to the expected number of common neighbors of nodes x and y in the configuration model [19].This paper use the abbreviation LHN1 to distinguish this index to another index LHN2 also proposed by Leicht, Holme and Newman.

g. Jaccard's Coefficient

Paul Jaccard introduces Jaccard coefficient over hundred years ago, which determines the association between two words. The Jaccard coefficient [20] is also known as Jaccard similarity coefficient. Jaccard index is a name often used for comparing distance, similarity and dissimilarity of the data set. To measure the Jaccard similarity coefficient between two data sets is defined as

$$j(x, y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \dots\dots\dots (7)$$

Jaccard distance is non-similar measurement between data sets. It can be determined by the inverse of the Jaccard coefficient which is obtained by removing the Jaccard similarity from (7). It is equal to a number of features that are all minus by number of features that are common to all divided by the number of features as presented below.

$$j\delta(A, B) = 1 - j(x, y) \dots\dots\dots (8)$$

This is the similarity of asymmetric binary attributes. Viewing the properties of an object in a binary format enables user to measure the similarity more easily by determining the objects A and B comprising "n" features.

The Jaccard similarity uses a measure of the share properties of both objects A and B whereas all of the Objects A and B given by 0 and 1 respectively.

h. Adamic Adar

N by N similarity matrix that contains the Adamic Adar similarity between every two nodes in the data sets. 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 node sand then defines the similarity between them. In case of link prediction in Social networks uses only topological information in which the features are Neighbours [7]. 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.

$$\text{Score}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \dots\dots\dots (9)$$

Finally this equation shows that the common neighbours of smaller degree more heavily.

i. Graph Distance

The graph distance between two vertices in a graph theory is determined by the number of edges in a shortest path connecting them. This is also called the geodesic distance**.** The distance function is a metric on the vertex set of a (weighted) graph G. The distance function satisfy this equation

$$d(a,b) \leq d(a,c) + d(c,b) \dots\dots\dots (10)$$

known as triangle inequality ∀ vertices *a, b, c* of G. This follows from the fact that if a person wants to get from *a* to *b*, then one possibility is to go via vertex *c*. Notice that there may be more than one shortest path between two vertices. If there is no path connecting the two vertices then conventionally the distance is defined as infinite. In the case of a directed graph, the distance d(u,v) between two vertices u and v is defined as the length of a shortest path from u to v consisting of arcs, provided at least one such path exists. Notice that, in contrast with the case of undirected graphs, d(u,v) does not necessarily coincide with (u,v), this may be possible due to one is defined while the other is not. If there is no path connecting the two vertices, i.e., if they belong to totally different connected parts then conventionally the space is outlined as infinite. In the case of a directed graph the space d(u,v) between 2 vertices u and v is outlined because the length of a shortest path from u to v consisting of arcs, provided a minimum of one such path exists.

j. Shortest Path

In graph theory, the shortest path problem is the problem of finding a path between two vertices or nodes in a graph such that the sum of the weights of its constituent edges is minimized. The shortest path problem can be defined for graphs whether these are directed, undirected or mixed.

$$S^{SP}_{x,y} = -d_{x,y} \dots\dots\dots (11)$$

k. Diameter and Radius

The two most commonly observed parameters in graph are radius and diameter. The diameter of a connected graph G is the maximum distance between two vertices is denoted as diam (G). The eccentricity of a vertex is the maximum distance from one vertex to any other vertex. The diameter is the maximum eccentricity among all vertices. The radius of a connected graph G is the minimum eccentricity among all vertices and it is denoted by rad (G). For a weighted undirected graph G

$$\text{rad}(G) \leq \text{diam}(G) \leq 2\,\text{rad}(G) \dots\dots\dots (12)$$

The upper bound follows from the triangle inequality, where c is a vertex of minimum eccentricity.

l. Hitting Time, Page rank and Variants

A random walk on G_{collab} starts at a node x and iteratively moves to a neighbour of x chosen uniformly at random. The hitting time Hx,y from x to y is the expected number of steps required for a random walk starting at x to reach y. Since the hitting time is not symmetric. It is also natural to consider the commute time.

$$\text{Score}(x,y) = Hx,y + Hy,x \dots\dots\dots (13)$$

These two measures used as score (x, y) can be defined as natural proximity measures. One difficulty with hitting time as a measure of proximity is that Hx,y is quite small whenever y is a node with a large stationary probability π_y, regardless of the identity of x**.** To counter balance this phenomenon, this paper tend to normalized versions of the hitting and commute times by defining

$$\text{Score}(x,y) = -Hx,y.\pi\_(y) \dots\dots\dots (14)$$

or

$$\text{Score}(x,y) = -(Hx,y.\pi\_y + Hy,x.\pi\_x)$$

Another difficulty with these measures is their sensitive dependence to parts of the graph far away from *x* and *y*, even when x and y are connected by very short paths. A way of counteracting this is to allow the random walk from x to y to periodically "reset," returning to *x* with a fixed probability α at each step where the distant parts of the graph will almost never be explored. Random resets form the basis of the Page rank measure for Web pages define score (x,y) under the rooted Page Rank measure to be the stationary probability of *y* in a random walk that returns to *x* with probability α each step, moving to a random neighbour with probability $1-\alpha$.

m. Clustering

One might seek to improve on the quality of a predictor by deleting the more "tenuous" edges in $Gcollab$through a clustering procedure, and then running the predictor on the resulting "cleaned-up" sub-graph. Consider a measure computing values for Score (x, y). This paper compute Score (u, v) for all edges in E_{old} and delete the (1-ρ) fraction of these edges for which the score is lowest and now re-compute score(x, y) for all pairs (x, y) on this sub-graph; in this way we determine node proximities using only edges for which the proximity measure itself has the most confidence.

B. Global Similarity Indices

a. Katz Index

Katz index [21] is based on the ensemble of all paths, which directly sums over the collection of paths and is exponentially damped by length to give the shorter paths more weights. The mathematical expression is defined as

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^{l}.|paths_{xy}^{<l>}| \text{ .....(15)}$$

Where$paths_{xy}^{<l>}$ is the set of all paths with length l connecting *x* and *y* and β is a free parameter (i.e., the damping factor) controlling the path weights. A very small β yields a measurement close to common neighbour, because the long paths contribute very little. The similarity matrix can be defined as

$$S^{Katz} = (I - \beta A) - 1 - I \dots\dots\dots (16)$$

Here β must be lower than the reciprocal of the largest eigen value of matrix A to ensure the convergence of Eq. 15.

b. Simrank

If two neighbours are so close to each other that they should be joined by an edge**.** Numerically, this is specified by defining similarity(x, x):=1 and

$$similarity(x,y) = \gamma.\frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} similarity(a,b)}{|\Gamma(x)|.|\Gamma(y)|} \dots\dots\dots (17)$$

for some $\gamma \in [0,1]$ is the decay factor. They finally stated score (x,y) := similarity(x,y). Simrank can also be interpreted by the random walk on the collaboration graph: it is the expected value of γ^{l}, where l is a random variable giving the time at which random walks started from *x* and *y* first meet.

VI. CONCLUSION

This paper is based on efficient social network which predict accurate relationship between links and measure different kinds of measurements for efficient link prediction. Link prediction is concern with the problem of predicting the existence of links among vertices in a social network. Link prediction techniques can provide very efficient way for discovering useful knowledge from available information. This survey is focus on modification of the existing methods to overcome their shortcomings or applying meta heuristic technique to improve accuracy of link prediction for easily find out the relationship between nodes.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition, ISBN 978-1-55860-901-3, 2006.

[2] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978-1-59904-252, Hershey, New York, 2007.

[3] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International

[4] Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September 2010.

[5] Nikita Jain, Vishal Srivastava ," Data Mining Techniques" vol: 02 Issue: 11 ,Nov 2013.

[6] Harsh J. Patel, Rakesh Prajapati, Prof. Mahesh Panchal, Dr. Monal J. Patel,"A Survey of Graph Pattern Mining Algorithm and Techniques" vol. 2, Issue 1, ISSN 2319 – 4847, January 2013.

[7] D. Sharma, U. Sharma, and Sunil Kumar Khatri, "An Experimental Comparison of the Link Prediction Techniques in Social Networks ",vol. 4.No. 1,February 2014.

[8] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of The social network of scientific collaboration. Physica A, 311(3-4):590-614, 2002.

[9] Lada. A. Adamic and Eytan Adar, "Predicting missing links via local information, Social Networks, vol. 25, no. 3, pp. 211-230, July 2003.

[10] 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.
[11] M. E. J. Newman, "Clustering & preferential attachment in growing networks," Physical review letters E, vol. 64, july 2001.
[12] M. E. J. Newman. The structure and function of complex networks. SIAM Review, 45:167-256,2003.
[13] M. E. J. Newman. "The structure of scientific collaboration networks". Proceedings of the National Academy of Sciences USA, 98:404-409, 2001.
[14] G. Jeh and J. Widom, "SimRank: A Measure of Structural Context Similarity," in Proc. the ACM SIGKDD International Conference onKnowledge Discovery and Data Mining, pp. 538-543, 2000.
[15] W. Liu and L. Lu, "Link prediction based on local random walk," Europhysics Letters, no. 5, 2010.
[16] Sourabh Vartak. A Survey on Link Prediction, University of New York Binghamton, NY – 13902, U.S.A. May 15, 2008.
[17] G. Kossinets, Effects of missing data in social networks, Social Networks 28 (2006) 247.
[18] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, MuGraw-Hill, Auckland, 1983.
[19] T. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1.
[20] E. A. Leicht, P. Holme, M. E. J. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2006) 026120.
[21] P. Jaccard, ´Etude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de la Societe Vaudoise des Science Naturelles vol.37, no. 547, 1901.
[22] L. Katz, A new status index derived from sociometric analysis, Psychmetrika 18 (1953) 39.