International Journal of Grid and Distributed Computing
Vol. 9, No. 2 (2016), pp.43-50
http://dx.doi.org/10.14257/ijgdc.2016.9.2.05

# A Review on Link Prediction in Social Network

Ajay Kumar Singh Kushwah[1] and Amit Kumar Manjhvar[2]

*[1,2]Department of CSE/IT*
*Madhav Institute of Technology and Science*
*Gwalior, Madhya Pradesh, India*
*[1]ajay.sati09@gmail.com, [2]amitkumar@mitsgwalior.i*

***Abstract***

*Social network analysis is an evolving field of research and link prediction problem shows a vital role for prediction of social network structure. This paper emphases on prevailing research on link prediction problem. Prevailing researches reveal that link prediction problem complexity, available solutions effective group communication management and social link consciousness. The link prediction problem across associated networks can include anchor link prediction problem and link transfer through associated heterogeneous networks. This paper summarizes recent growth about link prediction algorithms and survey of all the prevailing link prediction techniques.*

***Keywords****: Data mining, Link prediction, Social networkanalysis, Social network*

## 1. Introduction

The Link prediction problem is commonly described as a task to predict how possible a link survives between an arbitrary couple of nodes. In other words link prediction is the problem of recognizing whether a connection exists between two objects or not. Predicting variations to a social network is known as link prediction problem. The link prediction problem has been more properly defined as both the recognization of unnoticed 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 period t+1 given the state of a network at period t. As an example, consider a social network of co-authorship among researchers who are close in the network may be more likely to collaborate in the coming time. Link prediction is the merely sub-field of social network analysis, which has emphasis on edges between objects. Due to this reason, link prediction turn into more exciting than the traditional data mining areas which emphasis on objects. Link prediction can be used in many regions like recommender systems and criminal investigations.Approaches to link prediction have been anticipated based on various measures for analyzing the proximity of nodes in a network lead to the most accurate link predictions. In link prediction approach all methods assign a link weight score (x, y) two pairs of nodes x and y, based on given proximity measure and contribution graph G. A ranked list in reducing 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.

## 2. Background and Related Work

Data mining [1] refers to extracting knowledge from large data sets. The term data mining should have been more appropriately named as“Knowledge mining from data”
The comprehensive goal of data mining is to extract the useful information or knowledge from the stored data. In Data mining [2] there is an analysis of large quantities of data in order to discover meaningful patterns and rules. Data Mining [3] is about resolving

ISSN: 2005-4262 IJGDC
Copyright ⓒ 2016 SERSC

problems by analyzing data already present in the databases. Data mining tasks can becategorized into two categories descriptive and predictive. Descriptive mining tasks [4] focus on general properties of the data in the database. Predictive mining tasks focus on the current data in order to make predictions. The purpose of a data mining determination is normally either to produce a descriptive model or a predictive model.

Graphs [5] become important increasingly in modeling composite structures like circuits, chemical compounds, images and social networks. The graphrepresentation is basically used in pattern recognition and machine learning. Graph mining has become a key technique because of the increasing demand on the analysis of huge amounts of structured data in data mining.

A Social network [6] consists of a group of people and Links between them. These connections can be any type of social link that makes a relationship between two people. Social networks are popular way to mock-up the interactions among the people in a group or community. Social networks are highly vital in nature. They can grow and change as time variations and they can be visualized as graphs, in which a vertex denoted as a person in some group and link represents some form of association between the consequent persons [7].

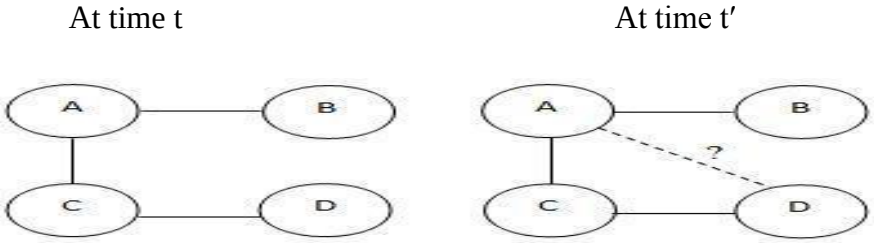


**Figure 1. Social Network Graphs G at Time t and t′**

Given in Figure 1 of the topology of a social network at period t, than it is need to predict the topology from period t to upcoming period t′ where t′>t. Assuming that the number of nodes does not change.

Lada. A. Adamic and Eytan Adar [8] anticipated the metric of similarity between two pages. It calculates the probability when two individual homepages are strongly related. It computes features that are shared among nodes and then defines the similarity involving them.

Liben-Nowell and Kleinberg [9] introduced a model based on node similarity for link prediction.There are numerous categories of node similarity. First one is the neighborhood based similarity like common neighbors of two nodes and the other one similarity based on a path which tries to resolve the shortest path distance concerning two nodes. So link prediction can be categorized into two classes, first is to the problem of identifying existing yet unknown links and predicting links that may come into sight in the future. M. E. J. Newman [10-12] used the concept of clustering & preferential attachment in rising networks. Glen Jeh and Jennifer Widom [13] anticipated the concept of Simrank. If two neighbors are so closer to each other than they should be connected by an edge.Liu and Lu [14] introduced a link prediction model based on the similarity of the nodes. This is significant in applications that consider the similarity of nodes such as gender, age etc.

## 3. Available Framework for Link Prediction

There are four dissimilar problems [15] given by link prediction are shown in the figure 2 below. The most of the research papers on link prediction spotlight on problem of link existence (whether a new link between two nodes in a social network will exist in the upcoming or not)**.** This is for the reason that the link existence problem can be easily

 Copyright ⓒ 2016 SERSC

prolonged to the other two problems of link load (links have different loads associated with them) and link cardinality (more than one link between the same couple of nodes in a social network) and the last problem of link type prediction is a little different which gives unlike roles to association between two objects. Classification whether a link exists or not can be achieved using a variation of classification algorithms like decision tree and support vector machine (SVM). Different structures like topological structures, content/semantic material of individual nodes can be used for analyzing the proximity of nodes in a social network.

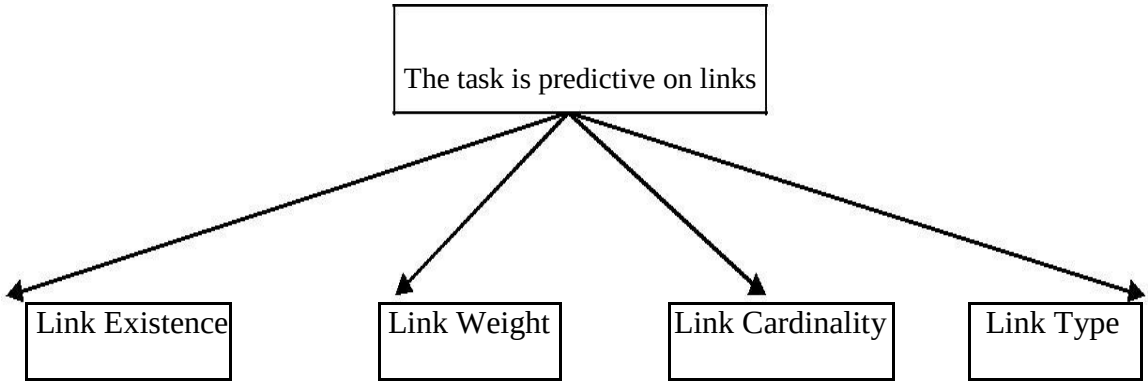


**Figure 2. Differentiation of Link Prediction Tasks**

## 4. Comparison of Basic Link Prediction Methods

**Table 1. Comparison of the Basic Link Prediction Methods with Respect to Function, Approaches and Running Time**

| Method | Function | Basic Implementation | Running Time |
|---|---|---|---|
| Shortest Path | $-dx,y$ | BFS | $O(V'.n^l)$ |
| Common Neighbors | $\mid \Gamma_x \cap \Gamma_y \mid$ | List Comparison | $(V'^2.nlogn)$ |
| Katz | $\sum_{l=1}^{\infty} \beta^l * \lvert paths_{xy}^{<l>} \rvert$ | DFS | $O(V'.n^l)$ |
| Simrank | $* \frac{\gamma \sum_{a\in \Gamma(x)} \sum_{b\in \Gamma(y)} similarity\ (a,b)}{\mid \Gamma(x) \mid\mid \Gamma(y) \mid}$ | Fixed point Iteration | $O(KV'^2.n^2)$ |

## 5. Measurements of Links for Prediction in Graph Mining

The structure of link prediction algorithm is based on the similarity of the nodes. Each couple of nodes x and y, is assigned a score $Sxy$. This function is defined as the similarity

Copyright ⓒ 2016 SERSC

between nodes x and y. Here introduces some straightforward link prediction similarity indices.

### 5.1. Local Similarity Indices

**5.1.1 Common Neighbors:** Common neighbor is a technique based on node neighborhood. The extent of common neighborhood of two nodes x and y can be defined as

$$S_{xy}^{CN} = | \Gamma_x \cap \Gamma_y | \quad (1)$$

Equation (1) represents the number of neighbors that x and you have in common. This technique is based on the intuition that if there is a node that is connected to x as well as you, then there is a high probability that vertex x be connected to vertex y. Thus, as the number of common neighbors grow higher, the probability that x and y have links between them increases.Kossinets and Watts [16] work to analyze a large-scale social network like Facebook. In their work, they suggest that two individuals having many common friends are very probable to be friend in the future.

**5.1.2. Salton Index:** Salton index [17] is defined as

$$S_{xy}^{Salton} = \frac{\Gamma_x \cap \Gamma_y}{\{\sqrt{Kx} * \sqrt{Ky}\}} \quad (2)$$

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

**5.1.3. Sorensen Index:** Sorensen Index is defined as

$$S_{xy}^{Sorensen} = 2 \frac{| \Gamma_x \cap \Gamma_y |}{kx+ky} \quad (3)$$

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

**5.1.4. Hub Promoted Index (HPI):** This index is offered for enumerating the topological overlap of pairs of substrates in metabolic networks, and is defined as

$$S_{xy}^{HPI} = \frac{| \Gamma_x \cap \Gamma_y |}{min\{kx,ky\}} \quad (4)$$

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.

**5.1.5. Hub Depressed Index (HDI):** Similar to the HPI, the HDI also considers a measurement with the opposite effect on hubs. It is defined as

$$S_{xy}^{HDI} = \frac{| \Gamma_x \cap \Gamma_y |}{max\{kx,ky\}} \quad (5)$$

**5.1.6. Leicht-Holme-Newman Index (LHN1):** This index consigns high similarity to node couples that have many common neighbors associated not to the possible maximum, but to the expected number of such neighbors**.** It is defined as

$$S_{xy}^{LHN1} = \frac{\Gamma_x \cap \Gamma_y}{\{kx * ky\}}. \quad (6)$$

Where the denominator $kx * ky$ is proportional to the likely number of common neighbors of nodes x and y in the configuration model [19].

 Copyright ⓒ 2016 SERSC

**5.1.7. Jaccard's Coefficient:** Paul Jaccard introduces Jaccard coefficient over hundred years ago, which is basically used to determine the association between two words. The Jaccard coefficient [20] is also known as the Jaccard similarity coefficient. Jaccard index is a name frequentlyrecycled for comparing distance, similarity and dissimilarity of the data set. To measure the Jaccard similarity coefficient between two data sets is defined as

$$J(x,y) = \frac{\Gamma_x \cap \Gamma_y}{\{\tau(x) * \tau(y)\}} \quad (7)$$

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

$$j\delta(A,B) = 1 - Jx,y \quad (8)$$

This is the similarity of a symmetric binary attributes.

**5.1.8. 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 likelihood when two particular homepages are strongly connected. It computes features that are common among nodes and then describes the similarity among them. For this major the features of the pages are calculated and then the similarities are defined.

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

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

$$S_{xy}^{SP} = -dx,y \quad (10)$$

**5.1.10. Clustering :** One might look forward to improve on the quality of a predictor by deleting the more unconvincing edges in $Gcollab$through a clustering procedure, and then running the predictor on the resulting cleaned-up sub-graph. Consider a measure, calculating values for Score $(x,y)$ for all pairs (x, y) on this sub-group.In this way we determine node proximities using only edges for which the proximity measure itself has the most self-assurance.

## 5.2. Global Similarity Indices

**5.2.1. Katz Index:** Katz index [20] is based on the joint of all paths, which straight sums over the collection of paths and is exponentially damped by length to provide the shorter paths further weights. The scientific expression is defined as

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l * \left|paths_{xy}^{<l>}\right| \quad (11)$$

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

$$S_{xy}^{Katz} = I - \beta A - 1 - l \quad (12)$$

Copyright ⓒ 2016 SERSC

Here β must be lower than the reciprocal of the largest Eigen value of the matrix A to ensure the convergence of Equation 12.

**5.2.2 Simrank:** If two neighbors 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

$$similarity\ (x.y) = \gamma * \frac{\sum_{a\in \Gamma(x)} \sum_{b\in \Gamma(y)} similarity\ (a,b)}{|\Gamma(x)||\Gamma(y)|} \quad (13)$$

For some γ∈ [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.

## 6. Conclusion

This paper is based on well-organized social network, which predict the exact association between links and measure unusual kinds of measurements for proficient link prediction. Link prediction is concerned with the problem of predicting the survival of links among vertices in a social network. Link prediction techniques can provide a very efficient way for discovering useful knowledge from existing information. This survey is the centerpiece on modification of the existing methods to overcome their shortcomings or applying meta heuristic technique to improve precision of link prediction for easily discover the relationship between nodes.

## References

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd edition, ISBN 978-1-55860-901-3, (**2006**).
[2] X. Zhu, I. Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978-1-59904-252, Hershey, New York, **(2007)**.
[3] J. Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining, vol. 1, no. 1, September **(2010)**. pp. 84-96,
[4] N.. Jain, V. Srivastava," Data Mining Techniques" vol. 02 Issue: 11, Nov **(2013).**
[5] H. J. Patel, R. Prajapati, Prof. M.Panchal, Dr. M. J. Patel,"A Survey of Graph Pattern Mining Algorithm and Techniques" vol. 2, Issue 1, ISSN 2319 – 4847, January **(2013)**.
[6] 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)**.
[7] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. "Evolution of The social Network of scientific collaboration".Physica A, vol. 311, no. 3-4, **(2002)**, pp. 590-614.
[8] A. Lada, "Adamic and Eytan Adar, "Predicting missing links via local information", Social Networks, vol. 25, no. 3 July **(2003)**. pp. 211- 230.
[9] D. L. Nowell and J. Kleinberg, "The link-prediction problem for social networking," Journal of the American Society for information science and Technology, vol. 58, no. 7, **(2007)**, pp. 1019 -1031.
[10] M. E. J. Newman, "Clustering & preferential attachment in growing networks," Physical review letters E, vol. 64, July **(2001)**.
[11] M. E. J. Newman, "The structure and function of complex networks",SIAM Review, 45:167-256,**(2003)**.
[12] M. E. J. Newman. "The structure of scientific collaboration networks". Proceedings of the National Academy of Sciences USA, 98:404-409, **(2001)**.
[13] G. Jeh and J. Widom, "SimRank: A Measure of Structural Context Similarity," in Proc. The ACM SIGKDD International Conference onKnowledge Discovery and Data Mining, **(2000)**, pp. 538-543.
[14] W. Liu and L. Lu, "Link prediction based on local random walk," Europhysics Letters, no. 5, **(2010)**.

[15] G. Kossinets, "Effects of missing data in social Networks", Social Networks, vol. 28 **(2006)**, p. 247.
[16] G. Salton, M. J. McGill, "Introduction to Modern Information Retrieval", MuGraw-Hill, Auckland,**(1983)**.
[17] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to the analyses of the vegetation on Danish commons", Biol. Skr. 5 **(1948)**.
[18] E. A. Leicht, P. Holme, M. E. J. Newman, "Vertex similarity in networks", Phys. Rev. E 73 **(2006)** 026120.
[19] P. Jaccard, ´Etude comparative de la distribution floraledarnsune portion des Alpes et des Jura, Bulletin

 Copyright ⓒ 2016 SERSC

de la SocieteVaudoise des Science Naturelles, vol. 37, no. 547, **(1901)**.
[20] L. Katz, "A new status index derived from sociometric analysis", Psychmetrika 18 **(1953)** 39.

## Authors

**Ajay Kumar Singh Kushwah**, He is currently pursing M.Tech. from Madhav Institute of Technology and Science, Gwalior, India. He completed his Bachelor Degree From SATI,Vidisha,MP, India. His Research interest includes Network Mining, Data Mining.

**Amit Kumar Manjhvar**, He is an Assistant Professor at Madhav Institute of Technology and Science, Gwalior, India. He has working experience of more than 5 years in different colleges. His research interest includes Network and Data Mining.

Copyright ⓒ 2016 SERSC

 Copyright ⓒ 2016 SERSC