Map-Reduce based Link Prediction for Large Scale Social Network

Ranjan Kumar Behera[1], Abhishek Sai Sukla[2], Sambit Mahapatra[3],Santanu Ku. Rath[4], Bibhudatta Sahoo[5], Swapan Bhattacharya[6]
Department of Computer Science and Engg.
National Institute of Technology, Rourkela, India.[1,2,3,4,5]
Jadavpur University, Kolkata, India.[6]
jranjanb.19@gmail.com[1], deva.abhi96@gmail.com[2], sambit9238@gmail.com[3],
rath.santanu@gmail.com[4], bdsahu@nitrkl.ac.in[5], bswapan2000@yahoo.co.in[6]

***Abstract*—Link prediction is an important research direction in the field of Social Network Analysis. The significance of this research area is crucial especially in the fields of network evolution analysis and recommender system in online social networks as well as e-commerce sites. This paper aims at predicting the hidden links that are likely to occur in near future. The possibility of formation of links is based on the similarity score between pair of nodes that are not yet connected in the social network. The similarity score, which we call link prediction score has been evaluated in Map-Reduce programming model. The proposed similarity score is based on both the structural information around the nodes and the degree of influence for neighboring nodes. The proposed algorithm is scalable in nature and performs quite well for large scale complex networks having good number of nodes and edges based on large pool of data or often termed as big-data. The efficiency and effectiveness of the algorithms are extensively tested and compared against traditional link prediction algorithms using three real world social network datasets.**

***Keywords*—Link Prediction; Preferential Attachment; Sim Rank; Jaccard Coefficient; Kartz Measure**

I. Introduction

Online Social Networks(OSNs) are an inevitable part of today's society. Research on this area has given rise to an entirely new field of research called Social Network Analysis (SNA). The size of social network is observed to be increase in a very large scale during a short span of time; as a result, it demands a necessity for analysis of such a large sized network. However, traditional tools are found to be bit inefficient in analyzing large scale network. Distributed computing framework may be considered as an alternative for analyzing such large network in reasonable amount of time. Understanding the evolution of social network is one of the important aspects of social network analysis, where the aspect of "link prediction" plays an important role. Predicting links which may come up in a network in future can be utilized in different applications, such as recommender system in e-commerce sites, OSNs and predicting hidden links in a terrorist network etc. But, large-sized networks, which are quite prevalent these days, give rise to a need for considering a highly scalable method, rather than using other conventional methods with higher time and space complexity.

The evaluation of link prediction may have certain difficulties. First one, sparsity of large complex networks may lead to difficulty in designing a statistical model due to prior existence of very few link. Secondly, it is a difficult task to develop a highly efficient algorithm for big real-time networks. Generally, there exists a trade-off between computational time complexity and accuracy, because as, we go on increasing accuracy, time-complexity increases and vice-versa. Hence, the challenge is to design an accurate and efficient algorithm for analysis of huge and sparse networks. This paper presents an approach for prediction of links considering mutual neighbors between two nodes, their shortest distance and their influence in the network. This algorithm has been implemented using Map-Reduce technique making it suitable for large-scale social networks. By considering the small-world effect and scale-free network, this algorithm is observed to have less implementation complexity.

The subsequent sections of the paper are organized as follows: In section 2, the related work in the field of link prediction in large scale social network is discussed. Motivation behind the Map-Reduce approach for link prediction has been presented in section 3. In Section 4, the proposed algorithm has been presented. Section 5 indicates its implementation part. In section 6, a comparative study is presented by using the graphical representation. Section 6 concludes the paper and presents the scope for future work.

II. Related works and Background Details

The problem of Link-Prediction has been a trending topic of research for the past few years in the field of SNA. The first promising work in this field was carried out by Libebn-Nowell and Kleinberg [4], where authors introduced the significance of topological information and discussed as to how it can prove to be highly effective in predicting links in social networks. It can be observed from their work that topological information is highly effective in comparison with picking up random edges from a social network, as these networks are sparse in nature. There have been a lot of works based on proximity evaluation for link prediction in the paper [1] These methods seem to be not that much effective in large-scale networks. For example, escape-probability concept has been proposed as a powerful measure of direction-aware proximity by Zhou and Jia [6], which is closely related to rooted page rank. But the proposed

DOI reference number: 10.18293/SEKE2017-100

TABLE I: Link Prediction Algorithms

Link Prediction Algorithm	Equations	Description
Common Neighbor [1]	$LP_{CN}(x,y) = \|N(x) \cap N(y)\|$	Probability of having hidden links between two nodes increases with the number of common neighbors.
Preferential Attachment [2]	$LP_{PA}(x,y) = \|N(x)\| * \|N(y)\|$	Hidden links are most likely to be observed between higher degree nodes rather then smaller degree nodes.
Adamic Adar [3]	$LP_{AA}(x,y) = \sum_{z \in \{N(x) \cap N(y)\}} \frac{1}{log(N(z))}$	Adamic adar is the measure that gives higher preference to those pair of nodes which have neighbors that are not shared with other nodes.
Kartz Measure [4]	$LP_{KZ}(x,y) = \sum_{i=1}^{\infty} \beta^i \left\|path_{x,y}^i\right\|$	This measure is defined as the sum of all the paths (less than diameter) between two nodes.
Sim Rank [5]	$LP_{SR}(x,y) = \gamma \frac{\sum_{a \in N(x)} \sum_{b \in N(y)} LP_{SR(a,b)}}{\|N(x)\|\|N(y)\|}$	Probability of establishing link between two nodes is more if they are connected with more similar neighbors.

method of computing escape-probability helps only to scale networks with large number of nodes [7].

Some of the conventional techniques used for prediction of links are listed in Table I. In all of these techniques, a score is assigned to every possible, but not yet connected, pair of nodes (x, y), based on the input snapshot 'G' of the social network at a certain time. Link Prediction heuristics predict links between nodes based on their similarity i.e., more similar the nodes, higher is their score.

III. Motivation

In a social network, it is observed that an influential node tends to involve in network evolution rapidly as, more and more nodes try to associate themselves with this node. This aspect has been the intuition behind algorithms like preferential attachment and rooted page rank [8], but it does not consider either the shortest distance between the concerned nodes or the number of mutual neighbors, which these nodes share. Similarly, in certain heuristics like Jaccard, Adamic-Adar etc., mutual neighbors between two nodes are being considered without considering the influences of concerned nodes. To put a check on these limitations, a method has been proposed in this study, which considers not only the popularity of nodes in a network, but also the geodesic distance and mutual neighbors between concerned nodes. The significance of mutual neighbors and influence of mutual nodes are observed to be the important parameters which are being applied in other link prediction algorithms. In a social network, if two persons share popular friends then, there is a higher possibility of them getting connected, than the case where mutual friends are not that much popular. The motivation behind the use of eigenvector centrality is that it provides centrality score, which considers the importance of its neighboring nodes.

IV. Proposed Algorithm

The Proposed algorithm for link prediction is based on the Map-Reduce programming model, presented in Algorithm 1. In this algorithm, probable score for link has been measured only for those pair of nodes between which no edge exists.

Algorithm 1: Link Prediction using Map-Reduce(LPMR) Model

Input: The large scale social network $G = (V, E)$ in edge list format and eigen-vector centrality for each node in vector form.

Output: Link predicted score between i and j where A[i,j]=0 i.e. node pair between which no edge exists

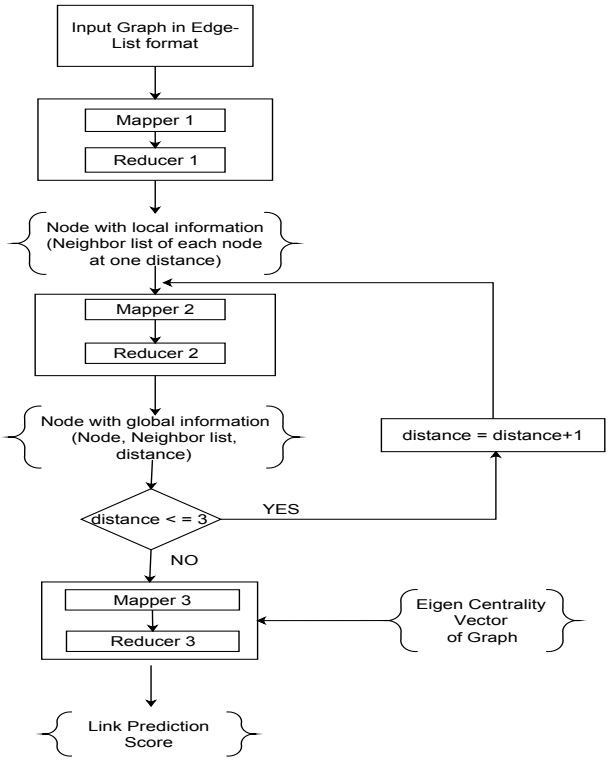


Fig. 1: Proposed Framework for Link Prediction

Step 1: The edge list is processed in first Map-Reduce phase. The first Map-Reduce phase is used to transform the edge list into adjacency list format, which contains local information of each node in the network. Local information may have list of neighboring nodes at one distance

Step 2: Modified adjacency list is then provided as input to the second Map-Reduce phase.

Step 3: Second Map-Reduce phase is used to apply extended ring search algorithm on modified adjacency list in order to get global information around the node.

Step 4: Global information contain structural information around a node upto path length four.

Step 5: Second Map-Reduce phase is repeated until the walk length is less than four. Neighbor list upto three walk length distance can be easily obtained in this step.

Step 6: The output of second Map-Reduce phase is then processed through Final Map-Reduce stage where link score

TABLE II: Real world datasets used for experiment

Datasets	No. of Nodes	No. of Edges	Diameter	Clustering Coefficient
Hamsterster Friendship [10]	1858	12534	14	0.901
Jazz Musician [11]	1133	2742	6	0.52
Ego Facebook [12]	2888	2981	9	0.0359

has been predicted.

Step 7: The link score has been predicted using the proposed similarity score which can be mathematically expressed as follow:

$$LS = \frac{1}{d}\left(EC(a) + EC(b)\right)\frac{\sum_{z\in\{N(a)\cap N(b)\}} EC(z)}{\sum_{k\in\{N(a)\cup N(b)\}} EC(k)} \quad (1)$$

where LS is the predicted link score between node a and b. EC(a) and EC(b) is the eigen-vector centrality score of node a and b respectively. d is the shortest distance between a and b.

The proposed LPMR model evaluates the similarity between nodes in the real world datasets. The initial input to the LPMR is the edgelist of graph G and eigen vector, where eigen vector centrality of all nodes has been provided. To compute shortest path of upto length L in graph G, customized form of expanding ring search algorithm is employed [9]. Traditional breadth first search (BFS) algorithm to find all shortest paths between any pairs of nodes has time complexity $O(n^3)$. Map-Reduce based algorithm has been applied for BFS in order to have global information of the node. Customized expanding ring search algorithm can find shortest paths of distance upto L from every node in a graph in a computational time complexity of $O(n)$, which is much less than the traditional algorithms. In this work execution time is further improved using Hadoop distributed frame work.

V. Implementation

A. Experimental Setup

Proposed Map-Reduce based Link prediction algorithms have been implemented on three real world social network datasets. All the experiments have been carried out on a cluster of 5 nodes, each with i7 processor having 3.4Ghz clock speed. Master node has the configuration with 1TB hard disk and 10GB RAM. It also acts as worker node. Each of other four nodes acts as slave or worker node. They all have symmetric configuration with 1TB hard disk and 20GB of RAM.

B. Dataset Used

The performance of the distributed link prediction algorithm i.e. LPMR is compared with the following existing link prediction algorithms:

a Common Neighbor (CN) [1]

b Kartz (KZ) Measure [4]

c Preferential Attachment (PA) [2]

d Adamic Adar(AA) [3]

e Sim Rank (SR) [5]

For testing our proposed algorithm, three real world social network datasets are considered. Details of the datasets are listed in TABLE II. The proposed algorithm use three pair of Map-Reduce components. Each component is intended for specific task of link prediction process. Input of the algorithm is the network with edge list format. As Map-Reduce takes only key-value pair as input, this format is to be strictly adhered in order to have compatibility with Map-Reduce model; however the dataset is transformed into a specific adjacency list format during different phases of Map-Reduce to gather the local information in the network. The first pair of Map-Reduce is used to discover the neighboring list of each node at one walk-length. After obtaining local information the dataset is then passed through second Map-Reduce phase where neighboring list of each node at four walk length has been discovered. This provides global information around the node in the network.

Link prediction score has been calculated in third Map-Reduce phase where eigen centrality of each node has been considered. Eigen centrality of each node has been provided at the beginning of the algorithm. The intuition behind using eigen centrality is that it captures the importance of each node by having connection with other important nodes. It can be observed that probability of establishing links between the node is high, if they are connected to more important nodes. The proposed link prediction score is also based on eigen centrality value of its common neighbors. It can be mathematically expressed as:

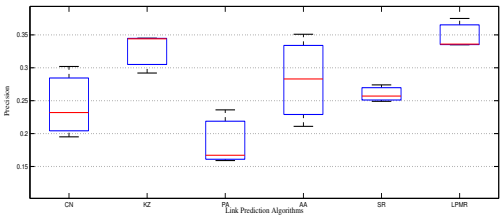
$$LS = \frac{1}{d}\left(EC(a) + EC(b)\right)\frac{\sum_{z\in\{N(a)\cap N(b)\}} EC(z)}{\sum_{k\in\{N(a)\cup N(b)\}} EC(k)} \quad (2)$$

where LS is the predicted link score between nodes a and b. EC(a) and EC(b) are the eigen-vector centrality scores of node a and b respectively. d is the shortest distance between a and b.

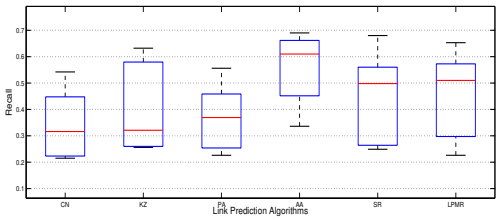
VI. Comparative Study

The performance of LPMR model is compared with other existing algorithms in terms of parameters such as precision, recall, f-measure and execution time by using three different datasets of social network. Few links have been randomly removed from the original datasets. After execution of link prediction algorithms, the predicted links have been compared with the links that have been removed prior to the experiments. All the experiments have been performed in Hadoop platform. The results have been aggregated and presented in the form of boxplot. Boxplot for Precision, recall, execution time and F-measure of link prediction probability is presented in Fig 2a, 2b, 2c and 2d respectively. Precision value for link prediction algorithms is presented in Table III. From Fig.2a, it can be observed that LPMR algorithm has better precision value as compared to those obtained using other traditional link prediction algorithms. Recall in information retrieval is the fraction of the links that are relevant to the hidden links being successfully retrieved. It is the number of detected links that actually exist before removal. From Fig. 2b, it can be identified that mean value of recall is better for Map-Reduce based algorithm.

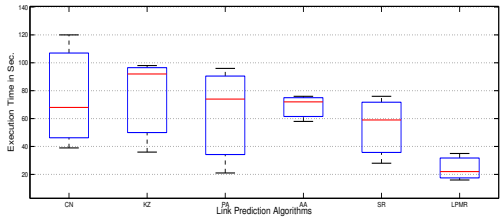
Execution time for the proposed algorithm i.e. LPMR is compared with other algorithms. The execution time (second)



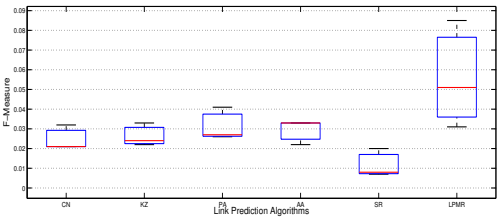
(a) Boxplot for Precision obtained in Different Algorithms



(b) Boxplot for Recall of Different Algorithms



(c) Boxplot for Execution time obtained in Different Algorithms



(d) Boxplot for F-Measure of Different Algorithms

Fig. 2: Comparative Analysis using Boxplot

TABLE III: Precision for different Algorithms

Network	CN	AA	Katz	SR	PA	LPMR
Hamsterster Friendship	0.232	0.211	0.292	0.249	0.167	0.385
Jazz Musician	0.195	0.283	0.345	0.257	0.159	0.336
Ego Facebook	0.302	0.351	0.344	0.033	0.274	0.375

TABLE IV: Execution Time for different Algorithms in Sec

Network	CN	AA	Katz	SR	PA	LPMR
Hamsterster Friendship	39	36	21	58	28	16
Jazz Musician	68	92	74	66	67	22
Ego Facebook	120	98	96	72	59	35

of different algorithms on different datasets are listed in Table IV. It can be observed from Fig. 2c that both minimum time and average time for the LPMR is comparatively less than other approaches. F-measure has been calculated for different algorithms on three datasets. F-measure value for Map-Reduce based algorithm is found to be more distributed as compared to other traditional algorithms. It can be observed from Fig.2d that mean value of F-Measure is more in LPMR algorithm .

VII. Conclusion and Future Work

Link prediction is one of the most important research directions in a number of application domains under SNA. It deals with revealing hidden links in the network. In this paper, an effort has been made in revealing hidden link in large scale network in distributed manner. Hadoop platform has been utilized for link prediction algorithm which is based on Map-Reduce programming model. The proposed algorithm has been extensively tested against few standard link prediction approaches. The Map-Reduce based Link prediction algorithm is found to be more suitable for processing large scale complex network. In Future, further enhancement to the proposed algorithm can be made for analyzing link prediction in dynamic network. Distributing processing tools like Storm and Spark can be implemented for streaming network where nodes and edges are added dynamically in continuous manner.

References

[1] Panpan Pei, Bo Liu, and Licheng Jiao. Link prediction in complex networks based on an information allocation index. *Physica A: Statistical Mechanics and its Applications*, 470:1–11, 2017.

[2] Benjamin Pachev and Benjamin Webb. Fast link prediction for large networks using spectral embedding. *arXiv preprint arXiv:1703.09693*, 2017.

[3] Eytan Adar and Lada A Adamic. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence*, pages 207–214. IEEE Computer Society, 2005.

[4] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[5] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[6] Wen Zhou and Yifan Jia. Predicting links based on knowledge dissemination in complex network. *Physica A: Statistical Mechanics and its Applications*, 471:561–568, 2017.

[7] Pulipati Srilatha and Ramakrishnan Manjula. Similarity index based link prediction algorithms in social networks: A survey. *Journal of Telecommunications and Information Technology*, (2):87, 2016.

[8] Erjia Yan and Raf Guns. Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, 8(2):295–309, 2014.

[9] Jahan Hassan and Sanjay Jha. Optimising expanding ring search for multi-hop wireless networks. In *Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE*, volume 2, pages 1061–1065. IEEE, 2004.

[10] Hamsterster friendships network dataset – KONECT, October 2016.

[11] Jazz musicians network dataset – KONECT, October 2016.

[12] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.