LINK PREDICTION

Report submitted in fulfillment of the requirements
for the Exploratory project of

Second year B.Tech

by

Shailendra Kori :21075080
Siddhant kharwar :21075083

Under the guidance of
Dr. Bhaskar Biswas Sir

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India

Declaration

We certify that

1. The work contained in this report is original and has been done by ourselves and the general supervision of our supervisor.

2. The work has not been submitted for any project.

3. Whenever we have used materials (data, theoretical analysis, results) from other sources, we have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever we have quoted written materials from other sources, we have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place : IIT (BHU) Varanasi

Date : 27 april,2023

Siddhant and Shailendra Kori

B.Tech Computer Science and Engineering

Indian institute of technology (BHU)

Varanasi ,Varanasi INDIA 221005

Certificate

This is to certify that the work contained in this report entitled "**Link Prediction**" being submitted by **Siddhant kharwar(Roll No. 21075083)** and **Shailendra Kori(Roll No. 21075080)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bonafide work of our supervision**.**

Place: IIT (BHU) Varanasi
Date: 27april,2023

Dr. Bhaskar Biswas
department of computer science and engineering
Indian institute of technology (BHU) Varanasi
Varanasi ,INDIA 221005

Acknowledgments

We would like to express my sincere gratitude to our professor Dr. Bhaskar Biswas Assistant Professor, Department of Computer Science and Engineering, Indian Institute of Technology (B.H.U.) Varanasi for guiding us at each and every step of our exploratory project. Without their assistance, it would have been an uphill task to arrive at the conclusions reached in this report.

Place: IIT (BHU) Varanasi
Date: April 27, 2023

Siddhant kharwar and Shailendra Kori

Abstract

Link prediction finds missing links (in static networks) or predicts the likelihood of future links (in dynamic networks). The latter definition is useful in network evolution **(Wang et al., 2011; Barabasi and Albert, 1999; Kleinberg, 2000; Leskovec et al., 2005; Zhang et al., 2015).** Link prediction is a fast-growing research area in both physics and computer science domain. There exists a wide range of link prediction techniques like similarity-based indices, probabilistic methods, dimensionality reduction approaches, etc. Similarity-based indices are extensively explored in this article. The experimental results of similarity are tabulated and discussed. We are going to explore some similarity based methods those are similarity based index and quasi local indices methods.

PROBLEM STATEMENT:

The problem statement for the exploratory project is that to develop accurate and efficient algorithms that can be used to predict a missing links in a network based dataset on the available information such as node attributes and edges.

MAIN OBJECTIVE OF LINK PREDICTION:

The primary object of the link prediction is to identify missing or future links in a network which can help in discovering hidden patterns and structures, understanding the network, and improving the network efficiency.

Link Prediction:

Link prediction is a task in network analysis that involves predicting the likelihood of a link or a edge between two nodes in a network. If given a network where some of the links are missing or not yet observed, the goal of link prediction is to predict which links are most likely to exist in the network.

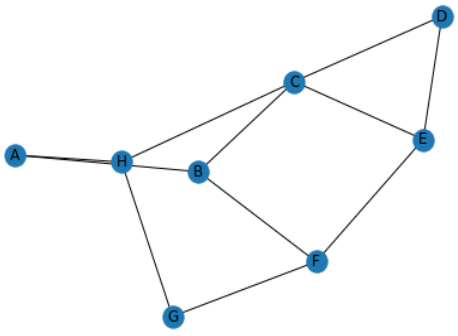


Figure 1: link between nodes.

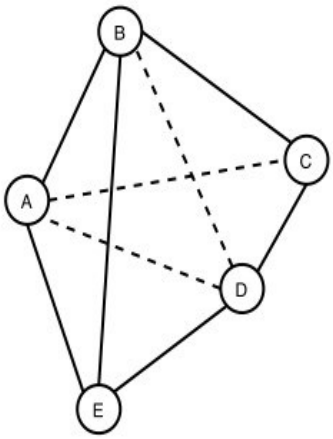


Figure 2. observed network

In figure 2,the links AB,AE,BC,BE,CD,DE are the links that are observed are visited. While the links AC,AD,BD are not observed. The link prediction will predict the links AC, AD, BD.

Link prediction can be used in various domains, including social network analysis, recommender systems(**like Facebook friend recommendation systems**), biology, and finance. The methods for link prediction vary depending on the type of network, the available data, and the specific application. Some common approaches include graph-based methods, similarity based methods, and machine learning based methods.

Project work

Methods used in the project :

Similarity-based methods:

In this project, I worked on the following indices for predicting the link:

1)Local-similarity indices:

Local indices are generally calculated using information about common neighbours and node degree. These indices give the chance of link being between the nodes. If the similarity is more between two nodes then there is high chance of link between those two nodes. Such indices are listed below:

i)Common Neighbours(CN):

The common neighbours simply count the number of common neighbours between two nodes.

$S(x, y) = |\Gamma(x) \cap \Gamma(y)|$, where $\Gamma(x)$ and $\Gamma(y)$ are neighbours of the node x and y respectively.

ii)Jackard Coefficient:

The Jackard coefficient is the ratio of the number of common neighbours between two nodes to the total number of unique neighbours of both nodes.

$S(x, y) = (|\Gamma(x) \cap \Gamma(y)|)/(|\Gamma(x) \cup \Gamma(y)|)$

iii)Adamic/Adar index:

The Adamic/Adar index is similar to the Jaccard coefficient, but it weighs the contribution of each common neighbour by the inverse logarithm of its degree.

Here, degree of a node refers to the number of edges incident on that node.

iv)Preferential Attachment(PA):

The preferential attachment assumes that the nodes with higher degree are more likely to form links.

v)Resource allocation index(RA):

The resource allocation index is similar to the Adamic/Adar index, but it weights the contribution of each common neighbour by the inverse of its degree.

vi)Cosine Similarity or Salton index:

The Salton index is the ratio of the number of common neighbours between two nodes to the square root of the product of their degrees.

vii)Sorensen Index:

Sorensen Index is similar to the Jaccard coefficient but it weights the contribution of each node equally.

viii)CAR based common neighbour index(CAR):

It is a modification of the Common Neighbours index, which considers the CAR index instead of the simple count of common neighbours.

ix)CAR based Adamic/Adar index(CAA):

It is a modification of the Adamic/Adar index, which considers the CAR index instead of the sum of the inverse logarithm of the degrees of common neighbours.

x)CAR based Resource Allocation index(CRA):

It is a modification of the Resource Allocation index, which considers the CAR index instead of the sum of the degrees of common neighbours.

xi)CAR based Preferential attachment(CPA):

It is a modification of the Preferential Attachment index, which considers the CAR index instead of the degree of the target node.

xii)Hub promoted Index(HPI):

The hub promoted index is similar to preferential attachment, but it gives a higher weight to the degree of the more connected node.

xiii)Hub Depressed Index(HDI):

The Hub Depressed Index is another variation of the Hub Promoted Index (HPI) that considers the impact of nodes with low scores on the network. It measures the ability of a node to reach other nodes in the network while considering the impact of its neighbouring nodes, as well as the impact of nodes with low HPI scores.

xiv) Local Naive Bayes-based Common Neighbours (LNBCN):

LNBCN is a measure of node similarity that uses a naive Bayes approach to compute the conditional probability of two nodes being connected, given their common neighbours. It takes into account the degree of the nodes, as well as the degree of their common neighbours, to calculate the probability of a connection.

xv) Leicht–Holme–Newman Local Index (LHNL):

The LHNL is a measure of the local clustering coefficient of a node in a network. It is calculated by counting the number of triangles that a node is part of, and then dividing this by the maximum number of triangles that could exist if all neighbours of the node were connected to each other.

2)Quasi local indices:

i)Local Path Index(LP):

Local Path Index (LP) is a graph-based index used in network analysis to quantify the local connectivity of nodes in a network. It measures the number of unique paths of length two between pairs of nodes that pass through a given node.

ii)Path of length 2(CH2_L2):

Path of length 2 (CH2_L2) refers to a sequence of two consecutive edges in a graph, where the first edge leads to an intermediate node and the second edge continues from that node to a destination node. In chemistry, this term may refer to a molecular fragment containing two carbon atoms connected by a single bond (also known as an ethylene group).

iii)Path of length 3(CH3_L3):

Path of length 3 (CH3_L3) refers to a sequence of three consecutive edges in a graph, where the first and third edges connect to the same nodes and the second edge connects to an intermediate node. In chemistry, this term may refer to a molecular fragment containing three carbon atoms connected in a linear chain (also known as a propyl group).

iv)Superposed Random Walk(SRW):

Superposed Random Walk (SRW) is a statistical method used to analyse complex networks. It involves generating random walks (i.e., sequences of nodes and edges) on a network, and then superimposing them to identify patterns of connectivity and clustering within the network. The method can be used to study various properties of networks, such as their degree distribution, clustering coefficient, and community structure.

Link Prediction using machine learning

First we calculated the several network features for each pair of nodes in a given dataset and creates a new data frame with these features and corresponding labels.

Calculated networks features are following:

Common Neighbors:

Jaccard Coefficient:

Resource Allocation Index

Adamic/Adar Index:

We define a binary value that indicates whether there is a link between the two nodes.

The resulting data frame can be used to explore the relationships between the different network features and the presence of links between nodes. For example, one could use this data frame to train a machine learning model to predict whether a link exists between two nodes based on their network features.

We used the **logistic regression model** on the new dataset that we have created and train the model to predict the link between nodes with the helps of level.

The features used for classification are all the columns except the label column then instance of the logistic regression algorithm is created using **Logistic Regression()**, and the model is trained on the training data using fit().

The model is evaluated on the testing data using the predict() method, and the accuracy of the model is computed using the accuracy_score() function.

The predicted probabilities are filtered using a threshold value, and only pairs with probabilities above this threshold are recommended as potential connections using **new_data[predicted_labels >= threshold][['From', 'To']]**. Finally, the recommended connections are printed .

Dataset:

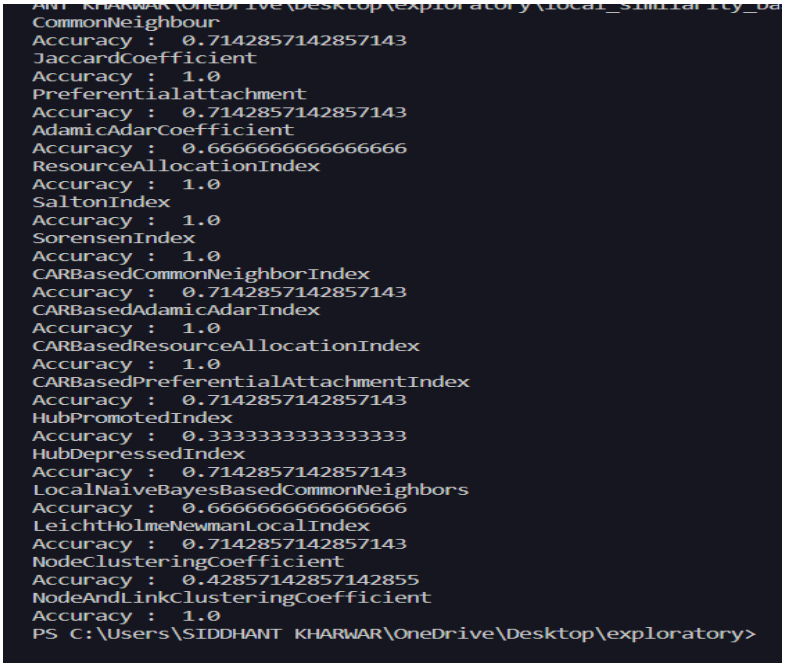
Karate.net

In this dataset the first row represents the number of vertices and number of edges in the graph. The remaining rows represent the edges of the graph.

RESULT:

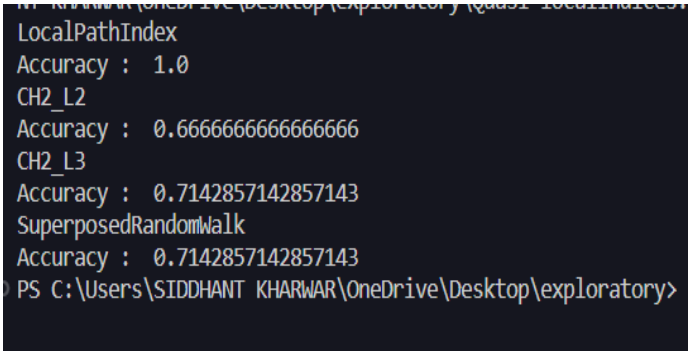
1) Local similarity based indices:

This is the accuracy of the different feature using the dataset karate.net.



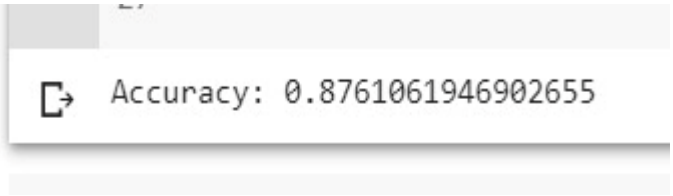
2)quasi local indices:

Result of the quasi local indices using the karate.net.



3) Logistic Regression:

This is the result of the model logistic regression using a binary label and karate.csv.



Conclusion:

Based on the values of accuracy given by the different indices for different dataset, we will be having following conclusions:
1. For dense networks with high clustering: Salton index may be useful as it is designed for dense networks.

2. For networks with highly skewed degree distribution: CAR based Common Neighbours or CAR based Adamic/Adar index may be effective as they take into account the degree distribution of nodes.

3. For networks with moderate degree of sparsity: Adamic/Adar index, Resource Allocation Index, or Jaccard coefficient may be useful as they are designed for networks with a moderate degree of sparsity.

4. For predicting links between highly connected nodes: Preferential Attachment or Hub Promoted index may be effective as they consider the degree of nodes.

5. For predicting links between nodes with low degrees: Adamic/Adar index, Resource Allocation Index, or Hub Depressed index may be useful.

6. For networks with a mixture of dense and sparse regions: Local Naïve Bayes Based Common Neighbours or Node and Link Clustering Coefficient may be useful.

So, we use different indices for different kinds of datasets for getting better accuracy. Some

indices may be better at some datasets and worst at some other datasets.

References:

Link prediction techniques, applications, and performance: A survey
Author: Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, Bhaskar Biswas