JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Study on Link Prediction in Networks: A Review

Avrodip Kumar Babai[1], Mr. Mukesh Sharma[2], Ms. Prashansa Taneja[3]

[1,2,3] Alakh Prakash Goyal Shimla University
Shoghi-Mehli By-Pass Road Shimla-171009
Himachal Pradesh, India
Email- [1]avrodip5658@gmail.com, [2]e0733@agu.edu.in, [3]Prashansa@agu.edu.in

Abstract: Link prediction in social networks is critical for mining and analysing the evolution of social networks because it predicts missing links in current infrastructure and digital or dissolution links in future networks. According to previous research, many similarity measurement techniques employ either node neighbourhood information or shared profile traits to calculate similarity between unconnected members in an online social network. There is no approach for link prediction that takes into account both node neighbourhood information and common profile features for calculating similarity. In this chapter, we discuss the specifics of our proposed link prediction approach, FINN (Feature Integration with Node Neighbor), which leverages neighborhood-based and attribute-based profile information to forecast missing connections.

Keywords: Link Prediction, Social Network, Similarity Metric, Dynamic Network, Learning Model

Introduction:

A social network is "a social structure consisting of a set of social actors, dyadic ties, and other social interactions among the social actors" [1] It represents the social framework among social entities. The social entities might be either a human or a business. Through linkages, social entities (or nodes) in online social networks are connected to other nodes (or edges). These linkages indicate the existence of a relationship or communication between the nodes. Figure 1 is an example of a social network in which the nodes represent social entities and the connections indicate the interactions between social entities. Utilization of online social networks has increased tremendously in the modern era. It has become a forum for millions of people to communicate their views, opinions, and ideas with one another. In addition to advertising, blogging, review gathering, and political awareness campaigns, this platform is also utilised for blogging. Online social networking websites and programmes such as Facebook, WhatsApp, Twitter, Flickr, and Instagram, among others, have become indispensable to our daily life.

Social networks are very dynamic, since they continue to expand (or evolve) over time. Increasing the number of links is the primary objective of a social network since it not only facilitates optimal use of the services supplied by social networks but also assures rapid transmission of information across the network. This is also clear from the goals of prominent social networking services such as Facebook, whose objective is to "bring the world closer together." LinkedIn, a well-known social networking tool, strives to "connect the world's professionals to increase their productivity and success." The establishment of new connections in social networks indicates the emergence of social interactions between individuals with shared interests, backgrounds, real-world contacts, etc. Understanding and assessing the factors that promote the growth of social networks is a complex matter. Examining the connections between two particular nodes helps simplify this problem. Link prediction refers to the topic of identifying the most likely connections between unconnected nodes in a social network. It evaluates the likelihood that a link may exist between two nodes that are currently unconnected. The likelihood is determined based on the presently existing nodes and links. It is conceivable for a user to know somebody in real life who he does not have a social network connection with.

Link prediction is an integral component of all social networking sites. Link prediction approaches try to create the maximum number of linkages feasible. Although it is impossible to have a completely linked network, it is always feasible to enhance the likelihood of connections between disconnected users.

On the basis of the existing network structure, link prediction algorithms in online social networks are applied to determine the future network structure. Existing link prediction systems use several similarity score heuristics to forecast new ties in social networks. These strategies are based on the triadic closure property, which implies that users who share connections have a propensity to be linked in the future. Some link prediction approaches rely on the neighbourhood of nodes, while others utilise shared profile characteristics. Existing link prediction approaches typically rely on one of two data modalities: node neighbourhood or shared profile traits. Existing node neighborhood-based link prediction approaches have the limitation of being able to anticipate just those connections that are directly related with neighbouring users. However, these algorithms are incapable of predicting relationships between users who are not directly related with their neighbours. For example, suppose there are four users A, B, C, and D, and they are linked as follows: A B, B C, and C D. If we want to anticipate the future connections of node A, then node neighbourhood-based link prediction is appropriate.

A link between disconnected node pair A and C can be foreseen, while a connection between disconnected node pair A and D cannot ever be predicted. This is owing to the fact that detached node pair A and D do not have a common neighbour. In contrast, profile feature based link prediction approaches employ the user's profile data to calculate the similarity score. These strategies are based on the homophily concept, which finds individuals with similar profiles. In a real-world setting, however, both data types are equally crucial for locating comparable people. In an ego network, there are four nodes from A to D. They are interconnected as follows: A B, B C, and B D. These nodes are linked to the profile characteristics "Workplace, Hometown, and Research Domain." The profile attributes of

nodes A, B, C, and D are "Delhi University, Delhi, Social Network," "Delhi, India," and "Social Network."

"Delhi University, Cornell, Bihar, Software Engineering," "Cornell, Maharashtra, Software Engineering," and "Stanford, Delhi, Social Network" A, C, and D are unknown to one another. Therefore, according to node neighbourhood-based approaches, a connection will be suggested between A C and A D, since they share a neighbour B. However, based on profile characteristics, A D will be more interested in forming a relationship than A C. This is because A and D have two profile characteristics (Delhi and Social Network). Thus, the combination of profile traits with neighbourhood data reduces the number of incorrect suggestions. In this chapter, we present FINN, a feature-integrated, node-neighborhood-based method to link prediction for online social networks. FINN combines the Jaccard coefficient and Adamic Adar Index to forecast probable future linkages between disconnected nodes in online social networks based on their network topology and shared profile characteristics. We assessed the suggested method, FINN, using a real-time Facebook network dataset obtained from the SNAP repository and confirmed the outcome using the area under the ROC curve as the assessment metric.

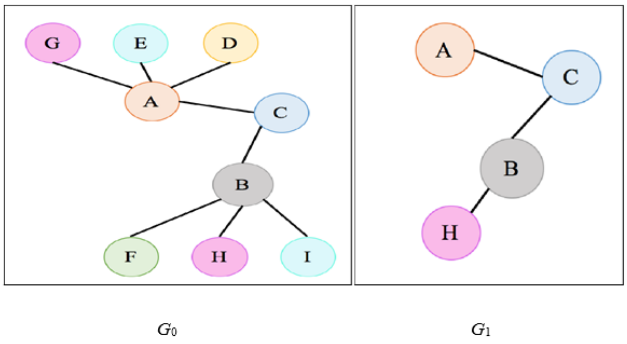

Figure 1: Hypothetical graphs, G_0 and G_1.

Literature review:

we explore current works that are closely connected to the work described in this dissertation. We split the effort associated with link prediction among several research streams.

score-based link prediction algorithms compute an accumulated similarity score based on either the neighbourhood of a node or on common shared attributes. It does not evaluate the connection pattern between connected pairs when predicting missing links. On the other hand, machine learning-based link prediction approaches predict a class label based on the connection pattern (in terms of number of shared neighbours or generally shared profile traits) between connected nodes. In online social networks, the issue of link prediction may be seen as a binary classification challenge for machine learning. Hasan et al. proposed machine learning classifications for missing link prediction in co-authorship networks to tackle the issue of link prediction. A collection of supervised machine learning classification methods are used to two distinct co-authorship networks in order to predict whether or not two writers will co-author a paper in the future. DBLP and Bio Base are the two co-authorship networks investigated in this

research. Hasan et al. examined multiple machine learning classifiers for link prediction, including Decision Tree, SVM, KNN, Nave Bayes, and Multilayer Perceptron, and discovered that all machine learning classifiers can handle the issue with acceptable accuracy. They did not, however, compare the effectiveness of machine learning classifiers with traditional similarity score heuristics of several types. Benchettara et al. In the E-Commerce network, a bipartite graph containing two kinds of nodes, users and goods, exists. A connection exists between a person and an item if the user has bought the item. The article described the use of a "indirect" characteristic for link prediction. A comparable item purchase count is utilised as a feature for all users who have bought a similar product. It has been shown that adding indirect features to the training set improves link prediction performance. O' Madadhain et al. used logistic regression to determine the likelihood of linkages between various node pairings. They used network information gathered from CiteSeer articles, AT&T phone calls, and Enron emails. Wang et al. Gong et al. used the support vector machine (SVM) technique to predict links on the Google+ website. Bliss et al. tackled the challenge of link prediction in dynamic social networks by training a classification decision tree. Pavlov & Ichise used several supervised learning algorithms, including SVM, Decision Tree, and J48, to forecast missing connections in co-authorship networks. They discovered that the performance of prediction varied based on the categorization approach used for link prediction. Deep belief networks (DBN) were used by Liu et al. to anticipate missing connections in signed networks. Each link in signed networks has a positive or negative sign. If two users agree or support each other, they have the potential to: There is a positive indicator on the border between them. Alternatively, if two users disagree with one other or oppose each other, a negative symbol appears between them. They discovered that deep belief networks may effectively forecast the establishment of links in signed networks.

Finn (feature integration with node neighbour):a new approach to link prediction for online social networks:

On the basis of the existing network structure, link prediction algorithms in online social networks are applied to determine the future network structure. Existing link prediction systems use several similarity score heuristics to forecast new ties in social networks. These strategies are based on the triadic closure property, which implies that users who share connections have a propensity to be linked in the future. Some link prediction approaches rely on the neighbourhood of nodes, while others utilise shared profile characteristics. Existing link prediction approaches typically rely on one of two data modalities: node neighbourhood or shared profile traits. Existing node neighborhood-based link prediction approaches have the limitation of being able to anticipate just those connections that are directly related with neighbouring users. However, these algorithms are incapable of predicting relationships between users who are not directly related with their neighbours. For example, suppose there are four users A, B, C, and D, and they are linked as follows: A B, B C, and C D. If we want to anticipate the future connections of node A, then node neighbourhood-based link prediction is appropriate.

A link between disconnected node pair A and C can be foreseen, while a connection between

disconnected node pair A and D cannot ever be predicted. This is owing to the fact that detached node pair A and D do not have a common neighbour. In contrast, profile feature based link prediction approaches employ the user's profile data to calculate the similarity score. These strategies are based on the homophily concept, which finds individuals with similar profiles. In a real-world setting, however, both data types are equally crucial for locating comparable people. In an ego network, there are four nodes from A to D. They are interconnected as follows: A B, B C, and B D. These nodes are linked to the profile characteristics "Workplace, Hometown, and Research Domain." The profile attributes of nodes A, B, C, and D are "Delhi University, Delhi, Social Network," "Delhi, India," and "Social Network."

Jaccard coefficient

The Jaccard Coefficient is a statistical approach for calculating set overlap. It is determined as the ratio of two sets' similarity to their diversity. It computes a score value ranging from 0 to 1. A greater similarity score indicates a greater likelihood that a link exists between two unconnected nodes. The score of Jaccard Coefficient is high when the number of common connections shared by disconnected nodes is large and their overall number of connections is low. Consider, for example, the graphs G0 and G1 in Figure 1. In graph G0, the unconnected pair of nodes A and B share a node C. The Jaccard Coefficient similarity heuristic yields a similarity score of 1/7 for the unconnected node pair A and B. Using the Jaccard Coefficient similarity heuristic, the similarity score for unconnected node pairs A and B in graph G1 is 1/2. This demonstrates that the Jaccard Co-efficient similarity heuristic is particularly advantageous for predicting relationships between users with fewer total connections. As the choice of threshold impacts the recommendations, the connections to additional nodes with a significant number of neighbours will also be projected. This attribute of the Jaccard Coefficient aids in achieving the objective of rapid information dispersion by facilitating the establishment of links between users who are separated and have few connections. If these individuals remain unconnected, the aim of rapid information dissemination in social networks is often thwarted.

Example of proposed approach on hypothetical net-work data:

This section provides an illustration of the suggested technique, FINN, using a fictitious network. Consider a hypothetical ego network with all three types of connections, i.e., "nodes with common neighbourhood and commonly shared profile features, nodes with only common neighbourhood but no commonly shared profile features, and nodes with no common neighbourhood but commonly shared profile features". In the supplied hypothetical ego network.

Node	Node Neighbourhood based Prediction	Feature based Prediction	FINN based Prediction
1	5	NIL	5
2	3,7, 8	3, 9	3,7,8,9
3	2, 7, 8	2, 9	2, 7,8, 9
4	NIL	6	6
5	1,6	NIL	1,6
6	5	4	4,5
7	2,3,8	NIL	2,3,8
8	2,3,7	NIL	2,3,7
9	NIL	2	2
TOTAL PREDICTIONS	**16**	**7**	**21**

Table 1: Prediction of newly created linkages by existing node neighborhood-based, profile-feature-based, and suggested technique, FINN on a hypothetical ego network with nine nodes.

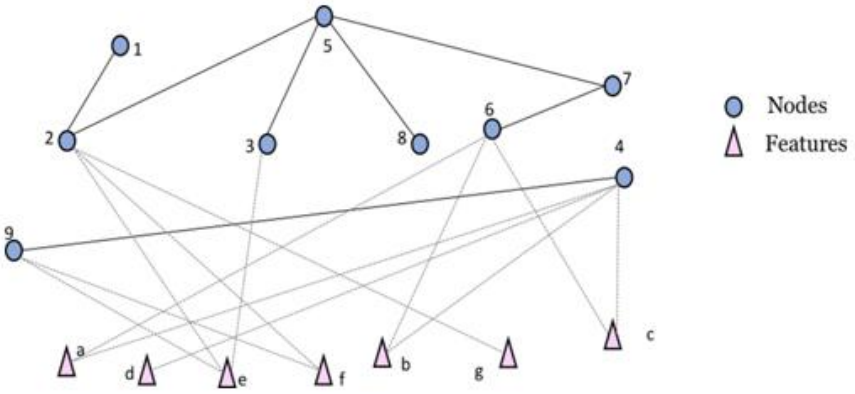

Figure 2: A hypothetical ego network with nine users and their profile features.

In the graph G, nodes are represented by dark circles and user profile characteristics are represented by solid triangles. The solid lines represent direct relationship between two nodes in the network, whereas the dashed lines represent profile characteristics of nodes. Nodes 2 and 3 in Figure 2 are unconnected and share a common node (5) and profile characteristic. Similarly, node pairs $1-5$, $2-3$, $2-7$, $2-8$, $3-7$ and $3-8$ are disconnected from each other but these node pairings are located in the same neighbourhood. The node pairings 2 9 and 4 6 have similar profile characteristics, however these node pairs do not share a common node connection. The common node neighbourhood based link prediction approach suggests a

connection between nodes 5 and 1 since they have a node (node 2) in common. Likewise, a link with nodes 3, 7, and 8 will be proposed to node 2 (because common node 5 is there), and so forth. Table 1 provides a summary of the results of implementing node neighborhood-based link prediction approaches. The node neighborhood-based link prediction approach will not offer any connection for node 4, since the node to which it is directly linked, node 9, has no additional connections. All predictions generated by connection prediction algorithms are bidirectional. If a link to node X is expected for node Y, then a link to node Y will also be projected for node X. Common profile feature-based link prediction approaches suggest persons based on similar profile characteristics.

Common profile feature-based link prediction techniques, for instance, will forecast the connections 2 3 and 2 9. This is owing to the fact that the unconnected node pairs 2 3 and 2 9 have a similar profile characteristic, e. Similarly, connections to nodes 2 and 9 will be anticipated for node 3 based on a shared profile characteristic, and node 4 will propose a link to node 6. (as a result of existence of commonly shared profile features, a, b and c).

Nevertheless, the suggested link prediction method, FINN, integrates the use of both node neighbourhood and common shared profile information of people to forecast likely ties in online social networks. For example, FINN predicts connections to nodes 3, 7, 8 and 9 for node 2 while node neighbourhood-based link prediction and common shared profile feature-based link prediction suggest connections to nodes 3, 7 and 8, respectively. Consequently, the suggested method, FINN, proposes all the available nodes that share neighbours, profile characteristics, or both. This will significantly boost the number of predictions generated for each user. To reduce the excessive number of forecasts, a threshold value is established. Only those node pairings with a similarity score over the threshold will be suggested to the user. The predictions produced by applying node neighborhood-based, feature-based, and FINN-based link prediction to the hypothetical ego network represented in Figure 2 are summarised in Table 1.

Experimental results:

To evaluate the prediction ability of the proposed FINN algorithm, experimental evaluations were done using the FINN algorithm to detect missing connections in different ego networks. Python programming and its accompanying libraries were used to extract the connection information and profile characteristics of each node. Following this, around 10 to 20 percent of the edges in each ego network were randomly deleted. The removed edges constitute the testing set. Eighty to ninety percent of the remaining connections represent the training set for our experimental assessment. While removing the test set's edges, we guaranteed the remaining network remained linked. Then, we used our suggested method, FINN, to identify the missing connections in the network (that we had initially eliminated from the training dataset). Figures 3 depict the experimental assessment outcomes of four ego network predictions. In Figures 3, k denotes the minimal number of profile traits shared by unconnected pairs of nodes within ego networks. The range of k's value is between three and six.

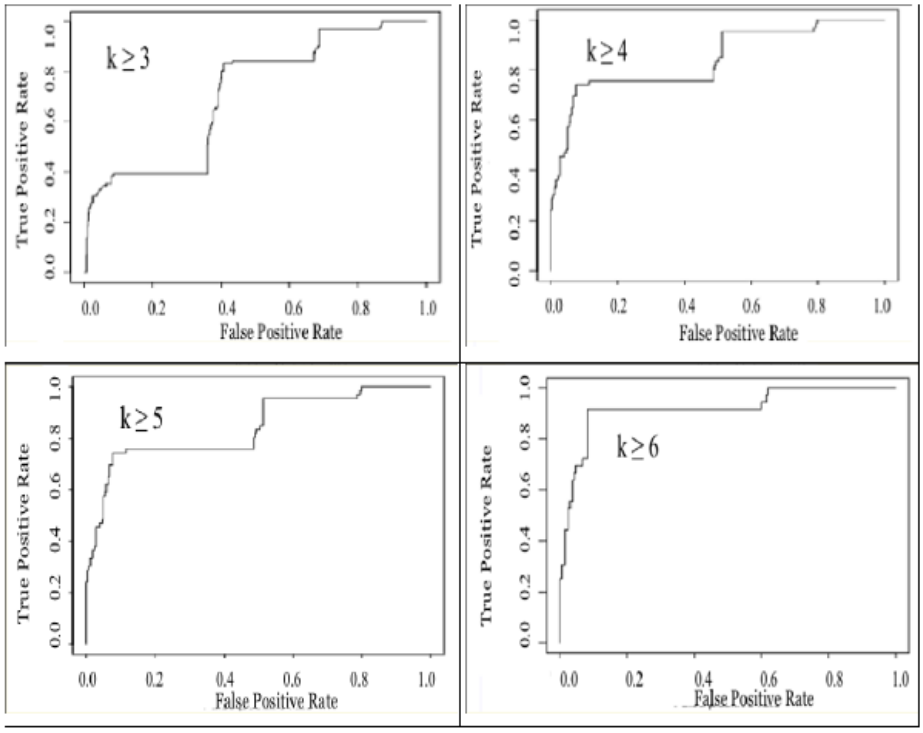

Figure 3: Evaluation results of our proposed approach, FINN on ego network 4 with 3, 4, 5 and 6 common features.

Conclusion:

Link prediction is essential for analysing online social networks, which have applications in several disciplines like data mining, bioinformatics, information retrieval, e-commerce, etc.. The need for rapid information distribution in various sectors has necessitated the development of link prediction algorithms. FINN is a feature-integrated node neighbourhood technique that has been suggested in this chapter. The suggested method, FINN, may be used to forecast missing links in social networks online. The motivation for the proposed method, FINN, is the notion that network structure and node properties are of equal importance for link prediction. FINN, the suggested method, computes a similarity score between pairs of disconnected users. The similarity score is computed by taking into account common shared node neighbours and common shared profile features between disconnected node pairs. For experimental assessment, we mined the dataset of Facebook network acquired from the SNAP repository. We tested the outcomes of the suggested strategy, FINN, using a well-known validation technique (Area under the ROC curve), and obtained up to 90 percent link prediction accuracy. Overall, we discovered that the relationships predicted by FINN have a good prediction performance when a large number of profile characteristics are shared.

References:

1. Backstrom L., Leskovec J., *"Supervised random walks: Predicting and recommending links in social networks"*, Proc. 4th ACM International Conference on Web Search and Data Mining, ACM, pp. 635- 644, 2011.

2. Krebs V. E., *"Mapping networks of terrorist cells"*, Connections, vol. 24, no. 3, pp. 43- 52, 2002.

3. Ressler S., *"Social network analysis as an approach to combat terrorism: Past, present, and future research"*, Homeland Security Affairs, vol. 2, no. 2, pp. 1- 10, 2006.

4. Airoldi E. M., Blei D. M., Fienberg S. E., Xing E. P., Jaakkola T., *"Mixed membership stochastic block models for relational data with application to protein- protein interac- tions"*, Proc. International Biometrics Society Annual Meeting, vol. 15, pp. 1- 34, 2006.

5. Zhu J., Hong J., Hughes J.G., *"Using markov models for web site link prediction"*, Proc.13th ACM Conference on Hypertext and Hypermedia, ACM, pp. 169- 170, 2002.

6. Li X., Chen H., *"Recommendation as link prediction: A graph kernel-based machine learning approach"*, Proc. 9th ACM/ IEEE-CS Joint Conference on Digital libraries, ACM, pp. 213- 216, 2009.

7. Talasu N., Jonnalagadda A., Pillai S. S. A., Rahul J., *"A link prediction based approach for recommendation systems"*, Proc. International Conference on Advances in Comput- ing, Communications and Informatics (ICACCI), IEEE, pp. 2059- 2062, 2017.

8. Kim M., Leskovec J., *"The network completion problem: Inferring missing nodes and edges in networks"*, Proc. SIAM International Conference on Data Mining, SIAM, pp. 47- 58, 2011.

9. Jahanbakhsh K., King V., Shoja G. C., *"Predicting missing contacts in mobile social networks"*, Pervasive and Mobile Computing, vol. 8, no. 5, pp. 698- 716, 2012