ISSN 2278-3091
Volume 10, No.1, January - February 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse271012021.pdf
https://doi.org/10.30534/ijatcse/2021/271012021

Link Prediction in Facebook using Web Scrapping and Deep Learning Techniques

Prof. Kalpana Prajapati[1], Dr. Harshal Shah[2], Prof. Rutvik Mehta[3]
[1]Parul University, India, kalpana.prajapati@paruluniversity.ac.in
[2]Parul University, India, harshal.mehta@paruluniversity.ac.in
[3]Parul University, India, rutvik.mehta@paruluniversity.ac.in

ABSTRACT

As Internet technologies develop continuously social networks are getting more popular day by day. People are connected with each other via virtual applications. Using the Link Prediction in social networks more people get connected, may be they are friends, may be work together at the same workplace and may be their education are. Machine learning techniques are used to analyze the link between the nodes of the network and also create a better link prediction model through deep learning. The objective of this research is to measure the performance using the different techniques to predict link between the social networks. Using deep learning, feature engineering can be reduced for link prediction. In this research, the feature based learning is used to predict the link for better performance. Dataset is obtained by scraping the profile of Facebook users and they are used along with the random forest and graph convolution neural network to measure the performance of link prediction in social networks.

Key words: Graph Convolution Network, Link Prediction, Random forest, Social network

1. INTRODUCTION

In recent days, link prediction is the current trend to analyze the social network and predict the future link for new upcoming friends in their list which are known or unknown. Many applications are used to predict the link such as friend recommendation, protein interactions, and e-commerce. The dynamic network is being changed over a period of time. So, dynamic networks are more challenging tasks than static networks. The Link predictions methods are calculated which is similar between nodes, find missing links and detect groups of objects and proximity of nodes. In early days, many efforts have been devoted to develop systems or tools to predict future link, based on historical data with state of the art methods which have some weakness like capacity and computational problems [14]. So to overcome these weaknesses various algorithms are designed in deep learning and try to predict better link prediction for the neighbor based networks. The algorithms are used to design for the undirected network or directed network. As per the Research, the supervised and unsupervised learning methods provide capability to improve link prediction results [22]. As an outcome, new interdisciplinary research directions have emerged in which social network analysis methods are applied to networks containing hundreds of millions of users. Miserably, missing link between users due to incomplete requirement of data or they are not cogitate in an online social network [8]. For the classification of links, they are necessary to extract the structural features in link prediction techniques which are the primary barrier. In this, it can be proposed a set of simple, easy computable structural features that can be analyzed to identify missing links. Using machine learning, it can identify missing links to classify links between users with one common friend [22].

In the proposed work using social network analysis various classification techniques are used to analyze the network, so the better performance can be analyzed. The real facebook network can be implemented using the scraping techniques to predict the link between users based on the features. The different classification techniques like random forest classifier graph convolution network are applied to predict the link and analyze the performance. So, the accuracy is compared with the different classification techniques.

2. RESEARCH SURVEY

In [1] link prediction is to find the missing link on a social network from the existing nodes and links. Link prediction is used for applications such as to find interactions between proteins, recommendation systems, security domain, co-authorship network. Here, paper presents a deep link to extract features for structural and content information. For the structural information, to consider path and community information and apply to classify WORD2VEC algorithm to find out the missing links in the given network for learning the deep link. For the content information, the whole information is extracted in one document using DOC2VEC algorithm. There may be links between two users based on their interests and learned features of topics in the network. After learning these two information they concatenate and apply deep learning algorithms like NODE2VEC, DeepWalk, LINE, M -NMF. Then after it compares AUC score using the different learning algorithm and neighbor based methods like common neighbor, jacquard coefficient, preferential attachment, Resource allocation, Adamic adar, Sorensen index. From this comparison Deep Link framework gives better performance for the limited resources. If the necessary structural and content information is used, then overall performance will be increased. So, the features are automatically extracted. For the future work, it can be adopted different deep learning techniques to improve the performance.

[2] Proposed network embedding algorithm to recommend friend for large heterogeneous network. It can be learned by edge embedding with multiple edge types into homogeneous components. The method splits the multi graph into a homogeneous sub-graph with one edge type. The deep walk and node2vec can be applied for the low dimensional space. Then, neural networks apply to train a heterogeneous edge embedding for link prediction. Over deep walk, Node2vec neural net improves 84% accuracy. The friend recommendation depends on this strategy and it conveys presently running at Hike network application.

In [3] link Prediction is the task to evaluate the existing link between the group of people or friends to find the future link between them. Many applications useful for link prediction like cyber security, biometrics, community structure, co-authorship structure. Link Prediction is a time dependent problem, where network changes over time snapshot of network graph. In this paper the future edge between two nodes can be predicted with accuracy using proximity feature vectors. Here it defines the structure of the link to add and remove edges dynamically. For the weak estimators, stochastic learning principles to apply update probabilities to new instances are observed with input vectors to construct a training set to train a deep learning network. With the sigmoid function in the output layer to bound the score between 0 and 1 for a single node. The datasets are Math overflow, EU-CORE, CollegeMsg used.

[4] Analyze the social networking websites allow social members to interact with each other positively as well as negatively. Such interaction is represented as a connection between them. The positive value means agreement of interactions in the graph and negative value means disagreement in the graph. Here in this paper, as to predict the view point of one user toward form evidence based on relationship with different users of the surrounding social network. Based on deep learning approaches an unsupervised method to represent features for link prediction and link prediction method by deep neural network based on Restricted Boltzmann machine (RBM). DBNS are learning a stack of RBM; there are three stack RBM used. Each of which is trained by using the hidden activities of the previous RBM as its training data. Hence, the proposed method improves the link prediction using features with joint distributions. Link Prediction in social networks is used to predict the future links which will occur in near future.

In paper [5] worked in a co-authorship network to analyze the features and metrics and assigned the score to all pairs of nodes then made an adjacency matrix. The co-authorship network node is an author and link shows two authors, which have written at least a paper together. The similarity based proximity scores are measured. After calculating score using feature vector and dataset of co-authorship data for all pairs of nodes implement the artificial neural network. From this paper, it also finds low scoring edges as weak links.

3. PROPOSED SYSTEM

Step 1: Facebook crawling using selenium tool
Step 2: Data extraction
Step 3: Prepare edgelist data
Step 4: Create network graph
Step 5: Apply proposed link prediction algorithm
Step 6: Training and testing using random forest and graph convolutional neural network classifiers
Step 7: Compare the result

4. METHODOLOGY DISCUSSION

Earlier approach used various supervised and unsupervised algorithm to achieve the task. It will be developed semi-supervised algorithm for scalable feature learning in networks. Using random forest classifier, there are different forest are used to analyze the network and classify the nodes to measure the performance. It can be optimized a custom graph-based objective function. In this, approach gives feature representations the maximum likelihood in a d-dimensional feature space to preserve network neighborhood nodes. In proposed model nodes in the network divided into different groups. It can be derived as K=(A, P) which has two parts, a method of partition A and a probability matrix P. let $P\alpha\beta$ be the probability of connected between two groups $\alpha\propto\beta$ and N be the observed network. So the probability of the network structure is p(N|K) = $P\alpha\beta l\alpha(1-P\alpha\beta)l\alpha\beta\alpha\leq\beta$. Here, $l\alpha\beta$ is the number of the existing link between $\alpha\propto\beta$ [6]. If there is a link exist then the predictor link is represent as a 1. And if there is no link then it become represent as a 0. The data are train and test using the splitting the nodes in random forest classifier and classify the links to gain the better performance. The graph convolutional network is used to analyze different datasets using deep learning. The graph structure is easily analyzed with the graph convolutional network to measure for better performance.

5. DATASET DESCRIPTION

In the Facebook dataset consists of user's profile parameter of about information like workplace and education. The features are automatically extracted by the selenium tool using crawling technique. Facebook application is not allowed to access user's features simultaneously, hence for the security reasons. So user can extract the feature using the slicing. It has to take time for the data gathering. The user is the node and features which are extracted those are edges. To prepare the larger dataset it may take the time complexity issue. The dataset prepare in the csv file of the all user's features here, in the dataset declare two features to make the edge list. In the future, there may be possibilities to add number of features to formation of edges of nodes. The number of edges formation is dependent on features matching with the user. Hence it can be said that if the number of features of the users increase the edges of nodes are also increased. Dataset prepare the edge list data from the facebook using the scraping technique. This edge list dataset is used to predict the link and analysis of network.

Dataset-1
Number of nodes: 408
Number of edges: 2595
Dataset-2
Number of nodes: 686
Number of edges: 6482
Dataset-3
Number of nodes: 978
Number of edges: 113695

6. IMPLEMENTATION PLANNING

For the data labeling preprocessing and policing of data using brute force as there is no specific algorithm for data labeling. To abstract the data set from Facebook using data crawling technique, which will give some unlabelled data, for such data a labeled data is needed for pre-prediction. The data set might have some missing data as well; it is needed to fill this missing data by predicting the output. Networkx package provides the graph data to create standard graph for reading in existing dataset, algorithms to analyze the resulting networks. For the graph the directed graph is converted from Digraph () and simple graph represent the undirected graph. To predict the missing data it will be used bagging with random forests algorithm. It will be decided the bagging method using the algorithm it will be filled the missing data and create a new data set with more features, then it will be converted the data into graph using NetworkX, For such it will be converted data set into nodes and vectors, Then it will be plotted the graphs for the data and see the connections between the nodes. The main aim of the research is to increase the accuracy of the existing system which uses single neural network for the link prediction. For link prediction, to do develop feature extraction algorithm that will follow the steps from start node to end node with the link value. For the graph structure it scans all links once. Then it stores the positive and negative link from to-to-from and from-to-to. It extracts the feature and save them in the graph. Here, the proposed system takes the features with different condition and compare the outcome with different classifier and model. Here the proposed system is used random forest and graph convolutional network to measure the accuracy.

7. OBSERVATIONS: RESULTS AND DISCUSSION

Here from the used methods for predicting the link, involved the use of the different deep learning techniques and algorithms such as convolutional neural network (CNN), deep Reinforcement learning, deep neural network and deep belief network . For the implementation of such methods various types of tools should be used. Then the data samples should be collected from the various social network websites for the implementation of such algorithms. In this approach, to predict the features from the raw dataset using multi label classification and after that the prediction link will be found using link prediction algorithm. Thus, finally compare the various techniques against the state of art methods of link prediction and achieve the better performance

Using the datasets prepared using the selenium testing tool from the facebook connected and unconnected profile features like workplace and education. Then from the dataset prepare the edgelist data and using the networkx the network graph is prepared. Then find the false positive rate because the dataset is small with the high degree. Where measures over the small dataset they give many false alarms. For the real datasets, there are two classification techniques are used one is the random forest classifier and graph convolutional neural network. Random forest technique is used for real datasets because of many forest network is present in network graph but it shows a less performance. Using the deep leaning the graph convolutional neural network classification technique is used to measure the better performance as compare to radom forest classifier. The link prediction methods of common neighbors, jaccard coefficient and adamic adar gives a same metrics as per the theoretical analysis.

8. COMPARISION OF CLASSIFICATION ALGORITHM

The study of the research implement with the real facebook dataset to measure the performance in terms of accuracy of link prediction using two classification algorithms. The random forest algorithm achieved the accuracy with the dataset-1 accuracy of 63% and the dataset-2 the accuracy of 61%. These results show that if the number of forest larger than the accuracy is increase. The dataset-1 and datset-2 proves that the number of nodes and edges size are not increase if the number of forest less. In dataset-3 one more feature adds in this dataset and experiments show that the forests are more than the dataset-1 and dataset-2. The accuracy has been improved. In the current research the graph convolutional neural network is used for the network analysis then it is implemented with the dataset-1, dataset-2 and dataset-3 the accuracy is improved then the above machine learning technique and previous research papers.

Table 1: Attributes of Cleveland dataset

Facebook Dataset	Random Forest	Graph Convolutional neural Network
Nodes :408 Edges:2595	63%	91%
Nodes : 686 Edges:6482	60%	95%
Nodes : 978 Edges: 113695	66%	98%

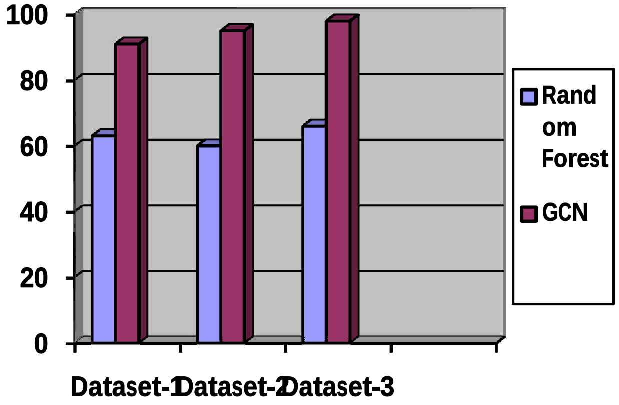

Figure 1: Comparison chart of different technique based on parameters of different dataset

9. CONCLUSION AND FUTURE SCOPE

Deep learning based algorithms are used to predict the future link and give better performance. There are some challenges in large scale networks to predict the future link and similarity between nodes. The various techniques are adopted and results are compared for link prediction. For the link prediction, here the parameter based learning technique is adopted. There are two classification algorithms; random forest and graph convolutional neural networks are used for link prediction. The feature based learning shows that the accuracy of graph convolutional neural networks gives a better performance as compared to other deep learning algorithms. The experiments have been done on real datasets. In future when adding new features, it can also give better performance to predict the link as compared to machine learning algorithms. It will be implemented with the various deep learning algorithms and performance can be measured with different accuracy parameters.

REFERENCES

1. M.Keikha, M.Rahgozar, M.Asadpour, **DeepLink: A Novel Link Prediction Framework based on Deep Learning**, 2018
2. J.Verma, S.Gupta, D.Mukherjee and T.Chakraborty, **Heterogeneous Edge Embedding for Friend Recommendation**, 2019
3. C.CHIU,J.ZHAN **Deep Learning for Link Prediction in Dynamic Networks Using Weak Estimators**, 2018
4. F.Liu, B.Liu, C.Sun, M.Liu and X.wang, **Deep Learning Approaches for Link Prediction in Social Network Services**, 2013
5. U.Sharma, B.Minocha, **Link Prediction in social networks: Similarity score based Neural Network Approach**, 2016
6. M.Lim , A.Abdullah , N.Jhanjhi ,M.Supramaniam **Hidden Link Prediction in Criminal Networks Using the Deep Reinforcement Learning Technique**, 2018
7. C.Zhang,H.Zhang,D.Yuan and M.Zhang, **Deep Learning Based Link Prediction with Social Pattern and External Attribute Knowledge in Bibliographic Network**, 2016
8. W.Xiong, T.Hoang and W.yangwang, **DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning**, 2017
9. Y.Meng,P.Wang,J.Xiao,X.Zhou, **NeLSTM: A New Model for Temporal Link Prediction in Social Networks**, 2019,IEEE
10. J.Chen,J. Zhang,X. Xu,D. Zhang,Q. Zhang, **E-LSTM-D: A Deep Learning Framework for Dynamic Network Link Prediction**, 2019,IEEE
11. T. Li ,B. Wang ,Y. Jiang ,Y. Zhang ,Y. Yan ,**Restricted Boltzmann Machine-Based Approaches for Link Prediction in Dynamic Networks**, 2018, IEEE Access
12. H.Mandal, M.Mirchev, S.Gramatikov, I.Mishkovski **Multilayer link prediction in online social networks, 2018**,IEEE
13. Y.Zhang, H.Wang **Link-Prediction and its Application in Online Social Networks**
14. X.Li, N.Du, H. Li, K.Li, J. Gao, A. Zhang **A deep learning approach to link prediction in dynamic networks**, 2014
15. W.Peng, X.BaoWen, W.Rong, Z.XiaoYu **Link Prediction in Social Networks: the State-of-the-Art**, 2014
16. D.John, L.Nowell **The Link Prediction Problem for Social Networks** , 2004
17. F. Aghabozorgi, M.R. Khayyambashi, **A new similarity measure for link prediction based on local structures in social networks** , 2018
18. W.Wang, L.Wu ,Y.Huang , H.Wang, R.Zhu, **Link Prediction Based on Deep Convolutional Neural Network**, 2019,MDPI
19. https://courses.cognitiveclass.ai/courses/coursev1%3ADeepLearning.TV%2BML0115EN%2Bv2.0/, 25/8//19, 13:05
20. https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff, [Access on 6/10/2019, 8:05]
21. https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/, [Access on 3/10/2019, 20:34] [22] M. AI Hasan , V.chaoji, S.Salem, M.Zaki, “Link prediction using supervised learning”, 2006
22. https://www.talentica.com/blogs/link-prediction-in-social-network/, ajeet shah, 22/11/2016