ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

A Comprehensive Study of Link Prediction Techniques for Stochastic Online Social Network

Shivshankar Rajput[1],Dr. Anil Rao Pimplapure[2]
(shivshankarrajpoot@gmail.com,pimplapureanil@gmail.com)
Eklavya University, Damoh(M.P.)

Abstract

In todays' era of machine learning and data mining the interpretation of data analytics has given a new paradigm to human beings. In this high computing era the data is interpreted as set of connected elements termed as network and for knowledge discovery that is for getting its instances and characteristics traditional machine learning and data mining are being focused. eventually this has emphasized over entity entity relations and perceptions now transformed from individuals to communities which has revealed a new set of problems related to the concept of network and this problems are not successfully solved. In other word considering an example where identification of communities of interconnected elements or entities, recognizing new or missing relations in between them or forecasting the role of any entity with in a community are some of the challenges.

Key-Words: Machine learning, data mining, social network, graph.

1. Introduction

The fundamental issues in social networks are about understanding the actual process which is responsible in evolution of social interaction between users or nodes. Out of these issues link prediction is just a part of that process. As discussed link prediction has commercial applications such as recommending friends on face-book and prediction or suggesting potential hires in professional network like linkdln. Perhaps considering the homophilly effect the exploitation of temporal link prediction where a sequence

of graph network G1….Gt from time 1 to t, how do it can predict link in further t+1. However link prediction and social networks can be defined in following section as it is necessary to understand the concept for discussing the topic.

1.1 Link Prediction

Link prediction is basically related with the detection of missing link between nodes on the basis of previous relations and links between nodes, prediction of links is also known as link prediction problem.The link prediction is a challenge in social networks where the prictioin of missing links which does not exist or perhaps exist but are unkown and having probability to occur in near future.

1.2 Online Social Network

A Social Structure referred as the collection of nodes (Individuals /Organizations/entities) and edges between these nodes referring the relationships on the basis of attributes. In order to understand this we can take an example where group of scientists, employees of a company, business leaders etc can be thought as nodes in a network and the probable coauthors of a paper, working on a project, serving altogether can be thought as edges respectively which is also represents a sort of association. The objective of Social Networks is to create

ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

opportunities to develop association, share information and enhance business in a network. Facebook and Twitter have are good example of social networking. The growth and dynamics of these networks has led several researches to study and examine the network properties such as structural and behavioral properties of large scale social networks.

According to the nature of social networks it is dynamic in nature can establish new relationships between nodes and breaking of many old relationships is a continuous process. These relational changes took place based on changes in characteristics of the nodes, characteristics of pairs of actors or link weights and random unexplained events that may influence the graph characteristics. The objective of analysis of network is to rank the edges or nodes. There are several ways to measure the ranking of nodes like degree centrality, closeness centrality, clustering co-efficient, Link prediction is prediction of the links that do not leave , and not known and have probability to occur in the near future. Identification of communities comprised of graph partitioning based on activities over the social network and determining the dense sub graphs in a social network. It is basically a network of social interactions and personal relationships between humans which for anlaysis we have represented by nodes and links amongst these nodes. Refer figure 1.1

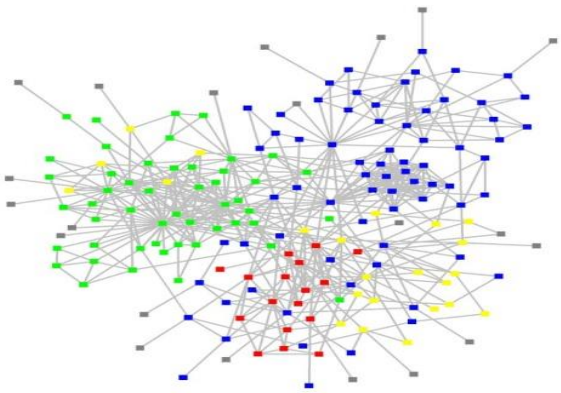

Figure 1 Social Network

1.3 The Link Prediction Problem

As we have discussed above that various kinds of edges are there between the nodes. Analyzing the social networks, at a given point of time there can be certain information about the relationships between the nodes that are not discovered or unknown. Link Prediction is the problem of identifying links that either don't yet exist at the given time t, but are unknown up to this time. Given a snapshot of a social network at time t, we need to predict accurately the links that will be added to the network during the interval from time t to a given future time t+1. Eventually the link prediction problem concentrates on up to…. what extent can the evolution of a social network to be modeled by using features of the network? Consider a co-authorship network among researchers, such that, there are various reasons, outside to the network, why two researchers who have never written a paper together will do so in the next few years. Either way when one of the researchers changes institution, they may come closer geographically. Such interactions are pretty hard to predict. But by studying the network characteristics, we could be in a position to predict the possible links that are going to form. The objective here in this is to make this problem exact that which link has to be identified, and to understand which measures of proximity in a graph that can lead for an accurate prediction.

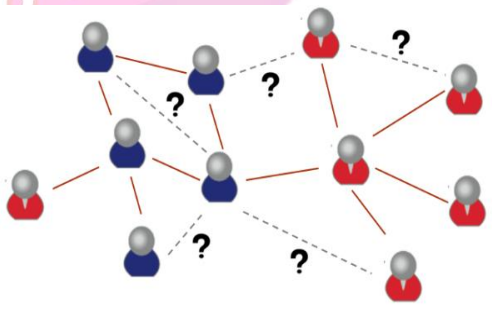

Figure 3: Link prediction

ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

2. Review of Literature

The literature review of existing researchers work done perhaps helps in converging the probable objectives of our research, thus this section of the synopsis briefly boast about the methods, techniques and models proposed by them.

According to D. Novell et. al. social network can be represented by G (V, E) where link represents a kind of interactions between its vertices on nodes at a given time t. consider snapshot of a social network at a given time. We select times t0 < t00 < t1 < t01, and give our algorithm to predict links to be formed in the future from the network G [t0; t00]. That results in predicting new links, which are not present in G[t0;t00], that are expected to appear in the network G[t1;t01]. It is refered to [t0; t00] as the training interval and [t1; t01] as the test interval [11].

Hasanet. al. in [12] has explored and analyzed social network they claimed that Link prediction is a key research area. In this research, they studied link prediction from the perspective of a supervised learning task with this they identified a set of features that can be superior in performance under the supervised learning establishment. The identified features are very easy to compute, and at the same time surprisingly effective in solving the link prediction problem. besides they explain the effectiveness of the features from their class density distribution. Then it was compared with different classes of supervised learning algorithms in terms of their prediction performance using various performance metrics, such as accuracy, precision-recall, F-values, squared error etc.and validated the results with set criteria also conducted certain experimental results on two practical social network datasets that shows well-known classification algorithms like decision tree, k-nn, multilayer perceptron, SVM which can predict link with surpassing performances, but SVM defeats all of them with narrow margin in all different performance measures.

Anticipating linkages among information items is a crucial information mining undertaking in different application spaces, including recommender frameworks, data recovery, programmed Web hyperlink era, record linkage, and correspondence observation. In numerous settings connect forecast is altogether in light of the linkage data itself (a noticeable illustration is the cooperative separating proposal). Connect structure based connection expectation is firmly identified with a parallel and practically isolate stream of research on topological demonstrating of huge scale diagrams. Diagram topological demonstrating expands on irregular chart hypothesis to discover closefisted chart era models imitating observational topological measures that abridge the worldwide structure of a diagram, for example, grouping coefficient, normal way length, and degree dissemination. These very much concentrated topological measures and chart era models have coordinate ramifications on connection expectation. This paper speaks to beginning endeavors to investigate the association between connection expectation and chart topology. The emphasis is solely on the prescient estimation of the grouping coefficient measure. The standard grouping coefficient measure is summed up to catch higher-arrange bunching inclinations. The proposed system comprises of a cycle arrangement interface likelihood display, a method for assessing model parameters in light of the summed up grouping coefficients, and model-based connection forecast era. Utilizing the Enron email dataset we show that the

ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

proposed cycle development display compared intimately with the real connection probabilities and the connection forecast calculation in view of this model beat existing calculations [1].

If we discuss about the ability to recognize the links amongst data objects is in center to various data mining jobs, like recommendation of a product and analysis of a social network. A reasonable literature has been available for the link prediction problem either perfectly embedded problem in specific applications or as a basic data mining task. The author here has a literature that mostly adopted a static graph where a network snapshot is analyzed as to predict hidden/future links. However, this representation is only appropriate to investigate whether a certain link will ever occur and does not apply to many applications for which the prediction of the repeated link occurrences are of primary interest. The author has introduced the time-series link prediction problem, considering temporal evolutions of link occurrences to predict link occurrence probabilities at a particular time. In line with the above experimental setup here also the author has used Enron e-mail data and high-energy particle physics literature co-authorship data, they have demonstrated that time-series models of single-link occurrences which achieve relative link prediction performance with commonly used static graph link prediction algorithms. Further they have also given a combination of static graph link prediction algorithms alongwith time-series model producing significant predictions over the static graph link predictions, which is demonstrating the great potential of hybrid methods exploiting both interlink that is spatial structural dependencies and intra-link that is temporal dependencies[2].

Apparently in a new research the authors have introduced a new way for link prediction in network structured domains, like in the Web, social networks, and biological networks. Here the approach is fully based on the topological characteristics of a network structures. A novel parameterized probabilistic model has been proposed by the authors to evaluate network evolution and has derived an efficient incremental learning algorithm for such models used to predict links in between nodes. In order to support this some promising experimental results have been established using biological network data sets [3].

For the traditional link prediction problem in a social network, snapshots of networks is used as a starting point just to recognize the further links, using graph-theoretic measures and the links that may appear in the near future. For this the authors has introduced a cold start link predictionmethod emerging as new approach for predicting the structure of a social network when it seems that the network itself is totally missing and only some of the information of the network nodes is available. They have proposed a two-phase method which is based upon the bootstrap probabilistic graph and the first phase perfectly creates a social network. Whereas in second phase probabilistic graph-based measures has been applied for the final prediction. The method is applied over a large data collected from Flickr, and experimental setups have confirmed the effectiveness of the approach[4].

The problems related to classification of objects and link prediction are extensively studied independently. Basically the classification of objects is done by assuming a complete set of known links whereas the link prediction is an assumption of a entire observed set of attribute nodes. Mostly in real world the attributes and links are often missing or incorrect and classification of

ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

objects is not given with the links related for correct classification and similarly prediction of link is not providing the labels required for correct link prediction. Keeping in consideration the authors has proposed a noval method addressing both the problems by intermixing object classification and link prediction collectively through an algorithm. They have investigated the conditions which improves the object classification and link prediction of any object in a network or over a wide range of network types [5].

Though the understanding about how individual mobility patterns shape and impact the social network is quite limited hence now it becomes necessary for the research community to develop a deeper knowledge and understanding of the network evolution and its dynamics. According to the authors this is partly unexplored as obtaining large scale society-wide data that perhaps capture the information on individual movements and social interaction is quite cumbersome.To this they have faced this challenge by tracking the projectile path and communication records of 6 million mobile phone users and found that movement of two individuals correlates their proximity with in the social network.

2.3 Parameters Used

The parameter used for computing future and missing links by various methods and techniques generally depends upon the diverse nature of the online social network. However we can broadly categorize them accordingly

1. Node/Vertex Features
    a) In degree
    b) Out degree
    c) All degree
    d) Node label
2. Edge/Link Features
    a) Edge label
    b) Edge weight
3. Distance or Path feature

2.3.1 Computation Techniques

As it is known that we have diverse nature of online networks available around, even in social networks and therefore so do the methods and techniques for link prediction in such networks. Above categorized parameters can be used in various computing techniques used for link prediction. However these techniques are generally local feature based and global feature based. Local feature based techniques focus on the node features and its neighbor neighborhood nodes whereas the global feature based techniques focus on overall path structure of the network. some of the classical techniques used under these are:

2.3.1.1 Local feature based

Common Neighbors (CN)

It is the simplest local technique where the similarity between two nodes is defined as the number of shared neighbor between both nodes [37]. It makes sense to assume that, if two individuals share many acquaintances, they are more likely to meet than two without common contacts.

Adamic Adar (AA)

It was basically proposed by LadaAdamic and Eytan Adar, where the intention was to measure the similarity between two nodes based on their shared features [38]. However, each feature weight is logarithmically penalized by its appearance frequency.

Jaccard Coefficient (JC)

This widely used coefficient in information retrieval systems was proposed by Paul Jaccard (1868–1944) to compare the similarity and diversity of sample sets [39].

ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

It measures the ratio of shared neighbors in the complete set of neighbors for two nodes.

2.3.1.2 Global Feature based

Katz Index

This index sums the influence of all possible paths between two pairs of nodes, incrementally penalizing paths by their length [40].

SimRank

SimRank is a method that computes how soon two random walkers starting from nodes x and y are expected to meet. This method is recommended for directed or mixed networks. After an exhaustive literature review it is found that there are various method or techniques which had been applied by research scientists to address the problem of link prediction. It has also been observed that there is no such method is available which can be generalized for all kinds of network due the the diverse nature of the online social networks, as we know that networks are dynamic in nature and developed according to the nature and requirement of the humans.

4. Scope of the study

This research proposal emphasize over link prediction problem specifically for directed social networks it is being understood that the task to mine a missing link can be termed as link prediction. Eventually the objective of link prediction is to recognize or guess new or missinglinks between two nodes of a social network as it has been mentioned that a social network can be represented as graphs where people are nodes and relationship between people is represented as edges which can again be termed as links. The prediction of these missing links would obviously guess on the analysis of previous observed relations that is layout of the social network (graph). We have gone through the research literature available addressing the problem of link prediction to recognize their present un-capabilities and limitations. It has been found thatlink prediction is now a days a powerful approach which can be applied over several real time social networking domains to predict futures trends enhancing an efficient forecasting system if its performance can be dealt accordingly. When talked about a social network then it is presumed and in-fact true that the network would be a large and continuously budding network comprising of millions of nodes and edges between them and analysis of and evaluation of node attributes and classifiying edge evolution in amongst them is cumbersome and needs high performance classifiers. Thus analyzing the social network particularly directed network and exploiting the node attributes and link prediction accordingly would basically the scope of this study.

5. Proposed Methodology

We will follow an algorithmic research methodology to complete this research work

1. The evolution of links prediction particularly from the perspective of directed graph along with an impact of machine learning and data mining techniques would be studied.
2. We would thoroughly go through the distinguished basics of evolution of social networks in order to classify them in directed and undirected graphs though our objective is restricted to prediction of links in directed graphs.
3. Machine learning and data mining schemes would be studied and applied over the available dataset and analyzed from the point of performance as it will help us in devising our method to predict missing links between the nodes by considering the diverse node attributes.

ISSN: 2455-6203
International Journal of Science Management & Engineering Research (IJSMER)
Volume: 09 | Issue: 03 | November - 2024 www.ejournal.rems.co.in
Date of Submission: 08/11/2024 Date of Acceptance: 15/11/2024 Date of Publish: 27/11/2024

4. A comparative study of exactly the available machine learning techniques applied for link prediction for identifying the missing links amongst the nodes would be done and subsequently a model would be proposed.

5. Developed model would be tested to validate and verify the performance using area under the ROC curve.

6. The results of proposed link prediction approach in directed social networks would be analyzed with the existing state of the art approaches for link prediction in the case of directed graphs.

6. Conclusion

The outcome of this research work will be a method for precdiction future links across the online stochastic social network. The network thus formed will help in identification of potential nodes which are contributing in evolution of the network.

References

[1] Z. Huang," Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient ", in LinkKDD'06, ACM 1-59593-446-6/06/0008, 2006.
[2] Z. Huang, Dennis K. J. Lin B,"The Time-Series Link Prediction Problem with Applications in Communication Surveillance", in INFORMS Journal on Computing, Vol. 21, No. 2, Springer 2009, pp. 286–303.
[3] H. Kashima, N. Abe," A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction",in ICDM '06 Proceedings of the Sixth International Conference on Data Mining, IEEE Computer Society Washington, DC, USA 2006, Pages 340-349.
[4] V. Leroy, B. BarlaCambazoglu, F. Bonchi. "Cold Start Link Prediction."The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Jul 2010, Washington DC, United States.12 p, 2010.
[5] M. Bilgic, G. M. Namata and L. Getoor, "Combining Collective Classification and Link Prediction," Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, NE, 2007, pp. 381-386.
[6] D. Wang, D. Pedreschi, C. Song, F. Giannotti1 A. L. Barabási "Human mobility, social ties and link prediction",Proceeding KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 1100-1108.
[7] J. Kunegis, A Lommatzsch, "Learning Spectral Graph Transformations for Link Prediction", ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning,Montreal, Quebec, Canada, June 14 - 18, 2009 ACM New York, NY, USA, pp. 561-568.
[8] Murata, Tsuyoshi, M, Sakiko"Link Prediction based on Structural Properties of Online Social Networks" H link prediction based on structural properties of online social networks" Springer, 2007.
[9] Md. Al Hasan, V. Chaoji, S. Salem, M. Zaki, "Link Prediction using Supervised Learning"
[10] LinyuanLü, T. Zhou.,"Role of Weak Ties in Link Prediction of Complex Networks",CNIKM '09 Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management, Hong Kong, China, ACM New York, NY, USA pp. 55-58, 2009.
[11] D. Nowell, Jon Kleinberg "The link prediction problem for social networks", CIKM,03 proceedings of the twelfth international conference on information and knowledge management, ACM, pp. 556-559, 2003.
[12] Al. Hasan, Mohammed J. Zaki, "A survey of Link Prediction in social networks", in Social Network data analysis, Springer March 2011, pp 243-275.
[13] A. Krishna Menon, C. Elkan, "Link prediction via matrix factorization", in ECML PKDD Machine learning and knowledge discovery in databases pp 437-452, LNCS Vol 6912 Springer, 2011.
[14] K.Y. Chiang, N Natrajan, A. Tiwari, "Exploiting Longer cycle for link prediction in signed network", CIKM '11 Proceedings of the 20th ACM international conference on Information and knowledge management Pages 1157-1162 Glasgow, Scotland, UK October 24 - 28, 2011 ACM New York, NY, USA,2011.
[15]. D. Wang, D Pedreschi, C Song, F. Giannotti, "Human mobility, social ties, and link prediction", Proceeding KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 1100-1108 San Diego, California, USA August 21 - 24, 2011.
[16] S. Scellato, A. Noulas, R. Lambiotte, C. Mascolo, "Socio-spatial properties of oline location-based social networks", Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.
[17] A. Narayanan, E. Shi and B. I. P. Rubinstein, "Link prediction by de-anonymization: How We Won the Kaggle Social Network Challenge," The 2011 International Joint Conference on Neural Networks, San Jose, CA, 2011, pp. 1825-1834.
[18] D. Liben-Nowell, J. Kleinberg,"The link-prediction problem for social networks", Journal of the American Society for Information Science and Technology 58, 1019–1031, 2007.
[19] Lada A. Adamic and Eytan Adar. "Friends and neighbors on the web.Social Networks" 25, 211–230, 3 2003.
[20] Paul Jaccard. 1901. ´Etude comparative de la distribution floraledansune portion des alpeset des jura. Bulletin de la Soci´et´eVaudoise des Sciences Naturelles 37 (1901), 547579.