Grenze International Journal of Engineering and Technology, Jan 2019

An Approach for Predicting Links in Social Network using Modified C4.5

Kaustubh Mahajan[1] and Prof. M. S. Bhamare[2]
[1-2]Maharashtra Institute of Technology, Pune, Department of Computer Engineering, Pune, India
Email: kstbhmahajan@gmail.com, mamta.bhamare@mitpune.edu.in

***Abstract*—Now days in the modern society online social network plays an vital role, link prediction is one of the most vital function on the online social network. There are variety of software of prediction of links which includes prescribing new things is available in market to users, friendship recommendation and finding spurious institutions. Prediction of link is a critical issue that provides to network evolution. The assignment is to evaluate the probability of the links of connections in future in a particular system. Social community relates humans through some kinds of connections. Therefore, it is all the more intriguing to predict the presence and the kind of a future link. This check the predicting the links in social networks because these latter allow concurrent institutions with a few types. On this gadget latent characteristic representation technique is used for representing the features from the consumer profile and KNN classification is used for link prediction. We archive parallellism using multithreading with parallel C 4.5 algorithm. This device will predict a friendship link on a social community on the premise of description or biography given inside the profile by means of the user.**

***Index Terms*— Friend recommendation, feature extraction, KNN classification, link prediction, Parallel C 4.5 algorithm.**

I. INTRODUCTION

In the network the working of link is to connecting among the two nodes. This basic idea can be utilized to present most complex frameworks where an extensive number of elements cooperate among them. The multiplication of information that can be presented as systems has made opportunities yet additionally new difficulties in the field of information mining. There number of researchers studied for in the problems of network mining some of the issues are community detection [8], analysis of structural network [9] and network visualization. A standout amongst the most fascinating system related issues is link prediction, which comprises of gathering the presence of new connections or still unknown interactions between sets of substances in light of their properties and the right links which are observed. The methodologies and procedures intended to tackle this issue empower the extraction of verifiable data show in the network and the identification proof of spurious connections, and displaying and assessing system development components. Techniques for link prediction have been effectively connected to biological networks with a specific end goal to foresee previously unknown communications between proteins, altogether lessening the expenses of exact methodologies.

They have likewise been utilized to show very powerful frameworks, for example, email or phone call

Grenze ID: 01.GIJET.5.1.5

systems. The techniques of link prediction are broadly present in our day by day lives, proposing individuals we may know however we are not in any case associated with in our social networks or items we could be keen on in electronic commerce [Huanget al. 2005]. Network have been widely considered since the recommendation of the principal fundamental models to recognize the laws that drive network development and prompt their basic features. Moreover it was not until particular link prediction arranged fundamental works, which played out a compressive investigation of the issue that this field went under the spotlight because of its pertinence and convenience in a great variety of settings.

Prediction of link is grounded in the observational proof that two elements will probably interface in the event that they are comparable. Comparability in networks must be comprehended as a theoretical idea and could differ between networks. Understand the area that the system represents is a vital advance to characterize the similarity between two nodes. In many spaces, it has been observed that nodes tend to shape exceedingly associated groups. This has prompted the normal definition of similitude as the measure of relevant immediate or indirect ways between nodes. One of the fundamental difficulty in connect forecast is accomplishing a decent harmony between the measure of data considered to perform the prediction and the algorithm complexity of the procedures expected to gather that data. Since genuine networks are normally framed by several thousand or even a huge number of nodes, the strategies used to perform link prediction must be exceptionally productive. In any case, considering just local data could prompt poor expectations, particularly in very sparse networks.

A. Prediction of links in social networks

Customarily, social network have been displayed in a single layer, where all nodes and edges are from a similar kind. The prediction of link issue considers the probability of presence of future link between two nodes in a network where there is no connection between them in the present state [9]. The link prediction problem has an extended records in modern facts sciences; be that as it may, it became first systematically presented for mining social network in [10].They figured the issue for a given social network G as to anticipate rundown of edges not introduced in G[d0, d'0], but predicted to exist in E[d1, d'1], where E[d, d'] indicates the sub graph of G at the time-stamp interim of [d, d']. [d0,d'0] is the training interim and [d1, d'1] is testing interim. Hasanet al. [11] expanded this model and demonstrated that one can get better outcomes by utilizing the graph topology, as well as logical data on the nodes and edges. Song et al.[12] utilized matrix factorization to appraise comparability between pair of nodes in a genuine social networks. Mining heterogeneous systems is more challenging than those with a similar sort of nodes and connections. Link expectation of heterogeneous systems has been contemplated in a few works [13]. Hristova et al.[12] communicated various research questions identified with the connection forecast issue in heterogeneous systems and proposed a model to tackle them. Meta paths were acquainted in with dissect heterogeneous systems, separate comparability, foresee links and rank their nodes.

Further paper is designed as in next chapter we will discuss about the different link prediction techniques, in next chapter we will discuss the proposed system with the experimental result and finally conclude the paper.

II. REVIEW OF LITERATURE

Shun YaoWu, QiZhang, MeiWu, [1], author represent complex systems as multi relational networks, and for extract a low dimension list all network latent model and find ratio test to examine correlation between factors.

Zong-wen Wei ,Rui Mao,Hao Liao, Mingyang Zhou,Yi- Cheng Zhang, Alexandre Vidmer, , [2], designed a hidden space mapping a network into Euclidean space based on the connection structures of a network. Which is compared with real geographical locations of nodes, and reconstructed locations are in conformity with those real ones. The distances between nodes in hidden space could serve as a novel similarity metric in link prediction. In addition, that hybrid the hidden space method with other state of- the-art similarity methods which substantially outperforms the existing methods on the prediction accuracy. Hence, the hidden space reconstruction model provides a fresh perspective to understand the network structure, which in particular casts a new light on link prediction.

Author Islam Elkabani and Roa A. Aboo Khachfeh, [3], studied on different technique to predict links these techniques are decision tree and the logistic regression approach. Here author predict the links over Facebook social network by using rough set theory depends on homophilic features. Some other techniques also used for link prediction.

Author Manfang Wu, Zhanquan Wang, HaiLong Hu [4] suggests some technique used to recommend the friends based on location. Here author proposed new approach to recommend.
Mingqiang Zhou, Rongchen Liu, Xin Zhao, Qingsheng Zhu [5] A multi-attribute classification method based on the order preference technique for similarity to the Ideal object (TOPSIS) was proposed to assess the local centrality of the common neighbouring nodes in an integral manner. To make local centrality TOPSI-based indicator achieve better results, a new weight calculation method for the attribute normalisation Matrix was also proposed. Experimental studies in 6 real world networks of disparate fields verified the superiority of the algorithm proposed in this work.
Shangrong Huang, Jian Zhang, Lei Wang, Xian-Sheng Hua [6], in this paper they investigate the structure of social networks and develop an algorithm for network recommendation based on the correlation of social friends (SFR). To achieve this goal, they correlate different "social role" networks, find their relationships and make friendly recommendations. The SFR based on NC is characterized by two key components: 1) the related networks are aligned by selecting important features of each network, and 2) the network structure should be preserved to the maximum before and after the alignment of the network. After an important selection of features has been made, they recommend friends based on these features. They carried out experiments on the network, Flickr, which contains more than ten thousand nodes and more than 30 thousand labels that cover half a million photos, to Show that the proposed algorithm recommends friends with more accuracy than the reference methods.
Asgari M, Jalili M, Orouskhani Y, Alipourfard N [7] studied link prediction problem in multiplex networks. Consider a multiplex network of Twitter and Foursquare now consider a social networks of the same users in these two platforms and developed a meta-path-based algorithm for predicting the links. To predict the links in Foursquare network The connectivity information of the two layers is used. For classifying the task Three classical classifiers are used.
Zhang Y Cho TW, Dave V, Song HH, [8], proposed two strategies for estimating proximity measure in OSN also for effectively and precisely evaluate the large family of proximity measures. Author also introduced the technique as incremental update protocol which enables real time proximity calculation in dynamic social networks.
Islam Elkabani and Roa A. Aboo Khachfeh [9] the author illustrate the use of rough set theory for predicting links for more Facebook-social network, which is based on the homophilic features. Other supervised learning algorithms are also employed in our experiments and compared to the rough set classifier, such as Naive Bayes, J48 decision tree, support vector machine, regression analysis, through and multilayer perceptron neural network. Moreover, they studied the influence of "common groups" and "total page likes" homophilic features in predicting friendship between users of Facebook, and also studied the effect of using the Jaccard coefficient for measuring the similarity of users homophilic attributes compared to using the overlap coefficient. They conducted experiments on two different datasets obtained from Facebook online social network, where users in each set, live at one and the same geographic area. The results showed that the rough set classifier significantly outperformed the other classifications for all experiments.

III. System Overview

A. Proposed System Overview

Detailed descriptions of the proposed system are as follows:

User Profile
Here system dataset of social network is used (example User face book profile is considered as an user profile).
System analyze the biography of the users of a social network websites such as weibo, faceboook and extract the data from it, also System extract general information of the users in the network

Pre-processing
Here Pre-processing is done on user Profile. System performs the TF-IDF operation. TF-IDF value of each data is calculated. In this system we compare the important words from the users profile.

Feature Extraction
Some Features are extracted from pre-processing. Latent feature representation (LFR) technique is used for feature representation. By Distributing users in a network latent feature of users are generated. Edge between

two users is generated through linking possibility which is defined in the latent-feature space. By using a maximum likelihood estimation method some parameters of existing users are extracted.

Classification

In existing system use naïve bayes and KNN classifiers for predict the link between the new users. We use C 4.5 algorithm for classification of users. Here we use parallel C 4.5 algorithm for achieving faster performance and better accuracy. Parallelism is induced in C4.5 with the help of multithreading concept wherein the information gain ratio is calculated in parallel environment. Considering the importance of the parallelism in any of the processes will reduce a lot of time for processing the data and finding results.

Predicting User

Now user Predicted Based on the result generated from C4.5 classifier and modified C4.5 which uses parallel processing.

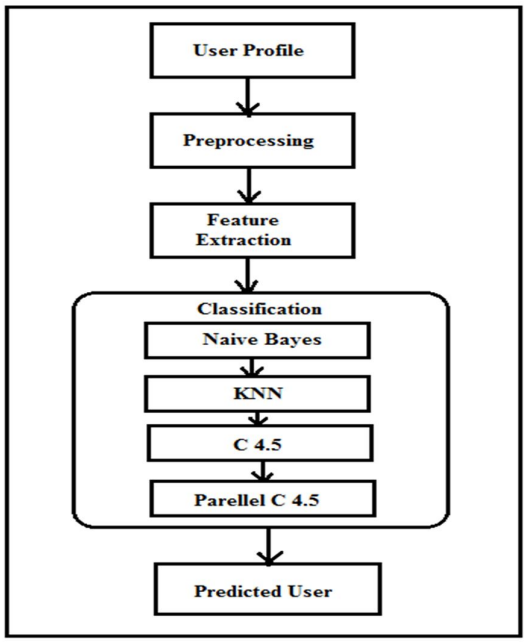

Figure .1 Proposed System Architecture

B. Mathematical Model

Let S be a system such that,
{S = {I, P, O, Sc, Fc}
where,
I = Input of the system
O = Output of the system
P = Processes in the system
Sc = Success Case of the output of the system
Fc = Failure Case of the output of the system
{I = fI1, I2,..,In}
{i = fUi, Dc}
where,
Ui = User Id of a user i
Dc = Description or Bio of the user
Process:
P = {P1, P2, P3}
P1: {Pr}
Pr = Preprocessing
Pr = {Lw, Po, Fr}

where,
Lw = Length of word

$$\frac{Len(word)}{Max\_Len}$$

Fr = Frequency

$$\frac{freq(word)}{\max\_freq} \times log_2 \frac{N+1}{n+1}$$

P2: {Lfr}
where,
Lfr = Latent Feature Representation
Lfr = {Vr, Pr}
where,
Vr = Vectors of binary latent feature of pairs of nodes i & j.
Vr = {Zi, Zj}
where,
$\overline{z_i}$== < $z_{1;1}$; :::::::; $z_{i;l}$; ::::::; $z_{i;L}$ > ($z_{i;l}$"f0; 1})
$\overline{z_j}$= < $z_{1;1}$; :::::::; $z_{j;l}$; ::::::; $z_{j;L}$ > ($z_{j;l}$"f0; 1})

Pr = The Probability of an edge Pij

$$P_{ij} = \pi_{l=1}^{L} \theta_l [z_i, l, z_j, l]$$

P3={Naïve Bayes}

$$P(x\backslash c) = \frac{P(x\backslash c)P(c)}{P(x)}$$

P4 = {Knn}
where,
Knn = {C^{wnn}}
Where,
C^{wnn} = Weighted nearest classifier
Similarity =$R_R(C_n^{wnn}) - R_R(C^{Bayes}) = (B_1 s_n^2 + B_2 s_n^2)\{1 + 0(1)\}$ for constant B1 and B2

$$s_n^2 = \sum_{i=1}^{n} w_{ni}^2$$

Where,
Ai & Bi are components of every vector.
P5={C 4.5}

$$Info\ (S) = -\sum_{i=1}^{k} \{[freq\ (C_i, S)/|S|] log_2 [freq(C_i, S)/|S|]\}$$

Where,
|S|= is the number of cases in the training set.
Ci= is a class, i=1,2,…k,
K= Number of cases.
Freq (Ci=S) and is the number of cases in Ci.
P6={Parallel C 4.5}
We archive parallism using multithreading with C 4.5 algorithm.
O = {O1, O2}
where,
O1 = Success Case Output.
O2 = Failure Case Output.

O1 = {Sc}
O2 = {Fc}

C. Algorithm

- Parallel C 4.5:

Step 1: Check the below base cases:
i) All the samples in the list belong to the same class. When this happens, all you have to do is create a leaf node in the decision tree that selects that class.
ii) None of the features provide any information gain. In this case, C4.5 uses the expected value of the class to create a decision node above the tree. All high nodes forgets traverse reshaped the sub tree of the node in the tree.
iii) If previously unseen class encountered then C4.5 creates a decision node using expected node.
Step 2: For each attribute, a thread is created and the gain ratio of the information normalized from the split on a is determined. Each thread runs independently, which improves performance.
Step 3: Make a best the attribute with the highest normalized information gain.
Step 4: Create decision node.
Step 5: Repeat the sub lists obtained by dividing by a best and add those nodes as children of the node.

IV. Results and Discussion

A. Experimental Setup Software required for the system is:

- JDK 1.8
- Netbeans Tool

B. Expected Result

Fig. 2 shows the time comparison between the existing and proposed system. From the graphs it is conclude that the time required for the proposed system is less than the time required for existing system.

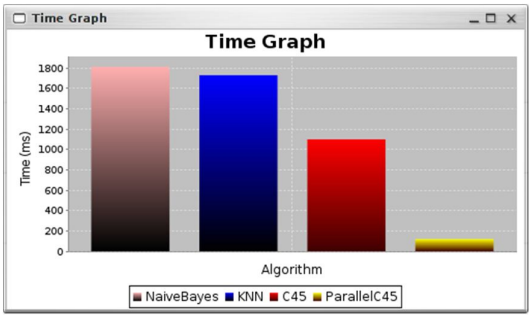

Figure 2. Time Graph

Fig. 3 represents the graphical comparison of memory consumption. It shows that the memory required for proposed system is less than existing system.

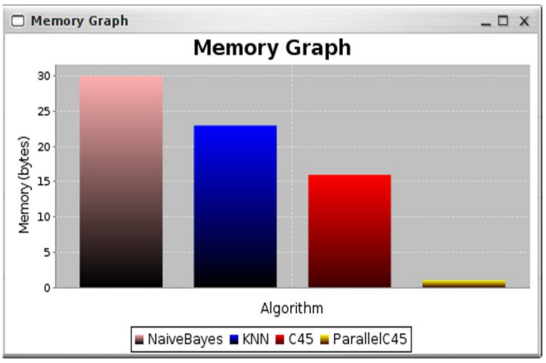

Figure 3. Memory graph

Fig. 4 shows the accuracy of the proposed system with the existing system. it shows that the proposed system is more accurate than existing system.

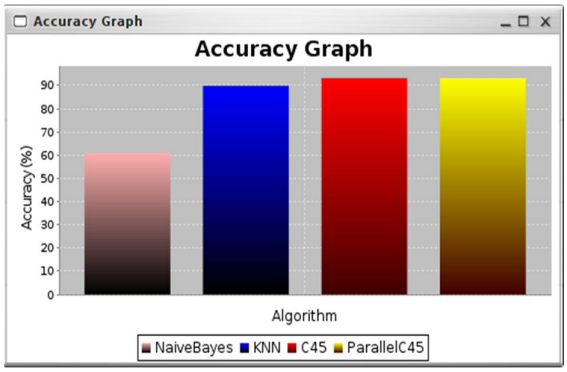

Figure 4. Accuracy graph

V. Conclusion

The study and design of a system for link prediction in social network sites is done in this work. So this study involves technology like social network analysis, link prediction, and feature representation. Different methods for the same approach were studied and found this approach to be more simple and feasible. Various data representation techniques were studied which are needed to represent the data which is to be fed to the link prediction algorithm. This is the first step towards designing a system which will predict links in social network to other nodes which behave similarly or have similar interests by means of description. The new aspect of this system is the use of the description provided by user in the profile. Using description, similar interests of users can be extracted using this system rather than using the fixed attributes like locations, workplaces, genders, schools colleges, etc. The future scope of this system also consist of using other attributes rather than description or bio data of user, in the system, such as dynamic location ,user preferences, network based locations, etc.

Acknowledgment

We should need to thank the examiners and furthermore wholesalers for making their advantages open. We also grateful to the reporter for their note worthy proposals moreover thank to the composition powers for giving the obliged base and support.

References

[1] Liang Duan, Charu Aggarwal, Shuai Ma, Tiejun Ma, Jinpeng Huai, "An Ensemble Approach to Link Prediction " IEEE 2017
[2] Zhiqiang Wang, Jiye Liang, Ru Li, Yuhua Qian, "An approach to Cold- Start Link Prediction: Establishing Connections Between Non-topological and Topological Information "IEEE 2016
[3] Sabrine Mallek, Imen Boukhris, Zied Elouedi, Eric Lefevre, "An Evidential Method for Multi-relational Link Prediction in Uncertain Social Networks " Springer 2016
[4] Fei Yu, Nan Che, Zhijun Li, Kai Li, Shouxu Jiang, "Friend Recommendation Considering Preference Coverage in Location-Based Social Networks" Springer 2018
[5] Mingqiang Zhou, Rongchen Liu, Xin Zhao, Qingsheng Zhu," Link Prediction Algorithm Based on Local Centrality of Common Neighbour Nodes Using Multi-Attribute Ranking " IEEE 2017
[6] Shangrong Huang, Jian Zhang, Lei Wang, Xian-Sheng Hua, "Social Friend Recommendation Based on Multiple Network Correlation " IEEE 2016
[7] Shun-yaoWu, QiZhang, MeiWu, "Cold-start link prediction in multirelational networks" in science direct, Volume 381, Issue 39, 17 October 2017.
[8] Yi-Cheng Zhang,Hao Liao1, Mingyang Zhou1, Zong-wen Wei1, Rui Mao1,Alexandre Vidmer1, "Hidden space reconstruction inspires link prediction in complex networks" March 2017
[9] Islam Elkabani and Roa A. Aboo Khachfeh, "Homophily-Based Link Prediction in The Facebook Online Social Network: A Rough Sets Approach" 2015.