International Journal of Engineering Research and General Science Volume 3, Issue 4, July-August, 2015
ISSN 2091-2730

Enhanced recommendation system using cluster based semantic link prediction

[1]BHAWNA [2]MEENAKSHI

1 Student, JMIT

2 A.P, Seth Jai Parkash Mukand Lal Institute of Engineering and Technology

Abstract- Links or more we say relationships, among data instances are everywhere. These links are considered as the patterns or relationships or similarity between users or objects. Most of the times, all the links of the social network are not present and we need to observe the non existing links. Main goal of the link prediction problem is to identify or predict the links that can occur in future. Link prediction problem has the major relevancy in social network. Traditional methods often gives less prediction accuracy as they are mostly structure based and calculates the user similarity considering simple terms like common neighbors. Although some methods use clustering techniques but there are some drawbacks and still much to improve. In our proposed method, we presented a cluster based link prediction (CBLP) algorithm that are time efficient and predicts link in more accurate way. In this method, Firstly k-means clustering technique is used to define the clusters but the no. of clusters are defined with the proposed formula; it leads to more efficient clustering of data occurs. Secondly, calculation of user similarity for all non existing links are there with the consideration of three proposed values such that a) ratio of number of common neighbours in the same cluster to that of total number of common neighbours b) semantic similarity value between the two nodes c) ratio of product of individual degree of two nodes to that of max degree of the network. With the proposed method, we are able to achieve more AUC value.

Keywords- Link Prediction, data mining, user similarity, social network.

1. Introduction

Most of datasets of interest today are best described as a linked collection of interrelated objects. The study of networks to predict the future links are of main concern. There are various techniques available to calculate the user similarity but there is much need to improve. Unlike, most of classical methods, proposed method works on cluster based link prediction in which the data is segmented into groups with the condition that the objects in a group are similar to each other and are dissimilar to the objects of other groups[1,3,14]. It is aimed that clustering is done in such a manner that intercluster similarity is minimum but the intracluster similarity is maximum. When the clusters are divided in an efficient manner then the work is to predict the new links or we say prediction of links that whether it will exist in future or not; and the prediction of the links are done with the consideration of the link strength between different users. Strength of the link is also named as score value between two nodes that checks the value upto which the two respective users are similar[2,8]. The learning paradigm of link prediction problem is to find out the most accurate similarity between pair of nodes. In social networks, there are issues that can be considered to generate the efficient user similarities. Therefore, the proposed method in this paper will put forward a new similarity calculation method which will consider the user similarity of the improved clusters with also the semantic similarity of different users. It is suppose to achieve higher accuracy rate in link prediction. Besides the own disadvantages, most of the cluster based methods lacks in some area[5,7]. As in case of k-means clustering, prediction of number of clusters that is k value is very difficult. If we are able to predict the good k value it by default generates good clustering of data. So in our proposed model we defined a formula to predict the number of clusters and clustering algorithm and also taking account of semantic similarity in addition to different clustering based similarity values; generating CBLP algorithm[3].

The remainder of the paper is organized as follows: In section 2, we review the related work or the background work related to link prediction. In section 3, we describe problem statement. In section 4, we define the proposed model. In section 5, we present experiment/performance evaluation. Finally In section 6, we provide the conclusion and future scope

2. Related Work

In this section we describe the methods that are proposed by different authors as the background related to link prediction problem

Fenhua Li,et al.[10] presents that the classical methods of link prediction are based on topological structure of graph and features of the path but very few of them consider clustering information. Actually, the clustering results contain the essential information for link prediction, and in these vertices common neighbors may play different roles depending on if they belong to the same cluster. Based on this assumption and characteristics of the common social networks, proposed a link prediction method based on clustering and global information. To satisfy the need of link prediction for social networks(i.e. simplicity and efficiency) and improve the accuracy of link prediction further, they proposed a new similarity measurement metric that combines the cluster information of nodes in networks and the topology structure information. These two features are considered while calculating the similarity metric or we say while providing score to the links. If first factor is multiplied by θ% then second is multiplied by 100-θ% and θ is a free parameter to select. With these factors; there is a improvement in accuracy of link prediction to 70%.

Jorge carlos valverde-Rebaza,et al.[11] proposed a new measure called WIC for link prediction between a pair of vertices in a network. In their model, three terms are considered that is common neighbors of intra cluster set or within cluster which is termed as W and common neighbors of between clusters or inter cluster which is termed as IC and the clustering technique to define the clusters. Clustering is a very useful concept for the predicting of future links so by using three different algorithm, WIC measure is proposed and it improves the accuracy of link prediction over local similarity measures.

Jungeun Kim,et al.[12] proposed LPCSP (Link Prediction inferred from Cluster Similarity and cluster Power) a novel link prediction method which exploits the generalized cluster information containing cluster relations and cluster evolution information that means static and temporal cluster information. In the static LPCSP uses cluster similarity and static cluster power defined by cluster's structure. LPCSP gives more weight when cluster similarity is higher and the structure of the cluster is more densely connected. In the temporal perspective, LPCSP gives more weight when the structure of the cluster is more strongly evolving. Cluster information consists of two major factors: (i) cluster similarity and (ii) cluster power. Two clusters are similar if there are many inter edges between them and cluster power based on both the static and temporal perspectives. The performance of LPCSP is best in four out of five datasets.

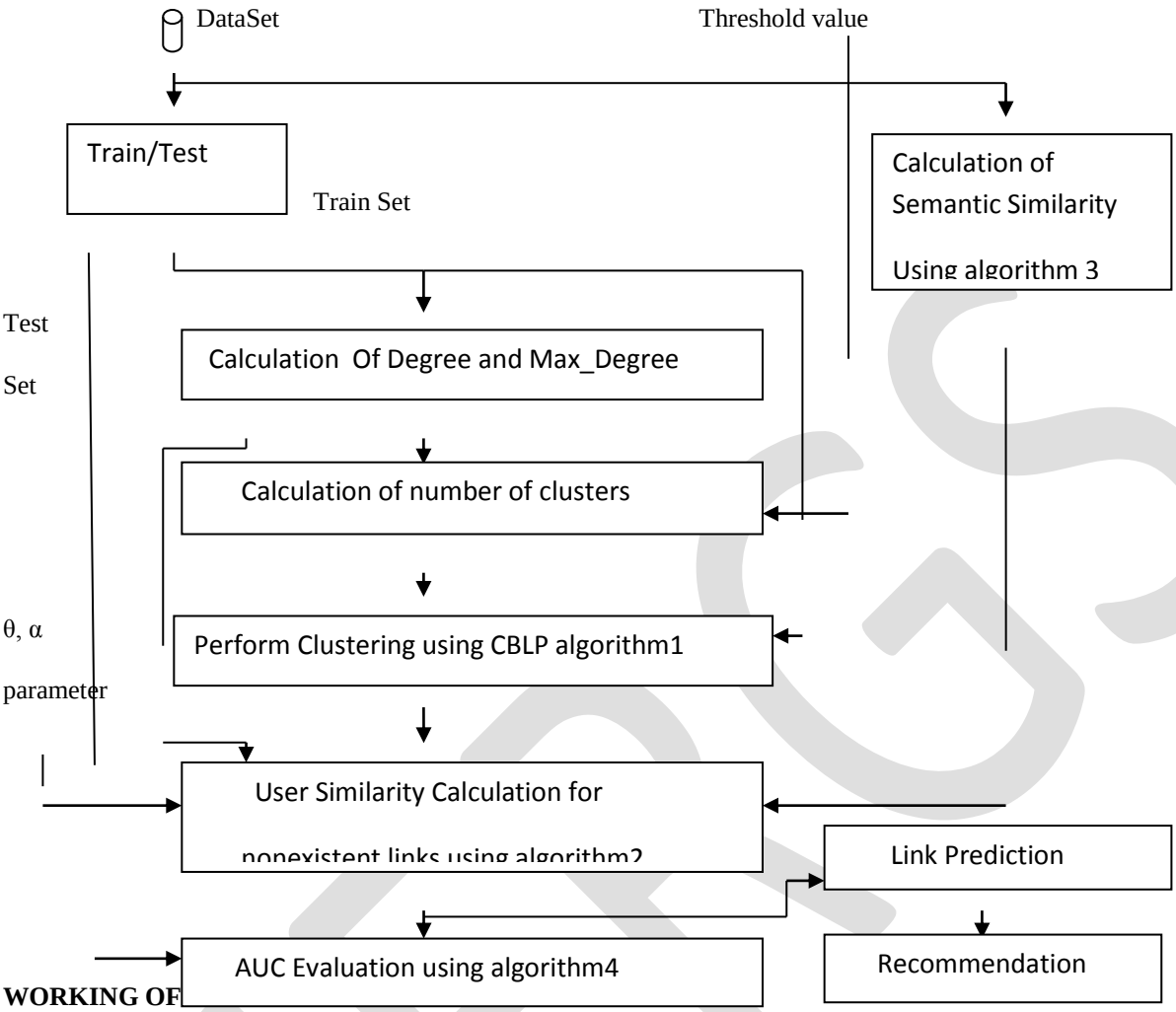
The above cluster based survey techniques and the classical methods without clustering have various disadvantages. Firstly, these methods shows less accuracy or we say the prediction value is less efficient and need much to improve. Secondly, classical methods that are only structure based have higher complexity and they do not consider clustering information. Our method not only improves the accuracy value but also provides flexibility as it has parameters that can be adjusted according to the requirements. The proof of achieving higher prediction accuracy will be stated in section 5 and section 6.

3. Problem Statement

In this section we will discuss about the problem statement that inspires us to work on the related model. As described in section 2, clustering information have relevant importance in predicting the non existing links. But with clustering also there are lots of issue that need to consider to improve the overall performance or the score of the user similarities. In the survey, it studied that there are lot of improvements are done on achieving the similarity between different users in a cluster or between two clusters. No doubt, with considering inter cluster factors and intra cluster factors it makes the technique rich but there are much need to make improvements over the generation of clusters. While performing the clustering there are issues that are need to improve Firstly, selection of number of clusters to be formed for a given dataset. Secondly, choosing the cluster to whom a user belongs perfectly. Thirdly, factors that need to give focus while calculating the user similarity. So, in our model we are aiming at considering the above mentioned issues to solve out. In section 4, we will describe the complete solution to the problem statement[10,3].

4. Proposed Work

In this section we describe the flow chart of the proposed model, then the complete description of the method with the various algorithms study of complete model.



WORKING OF

Step1 Train/Test Splitter- It is used to split the dataset into training data and testing data. Training of the data must be done in such a manner that atleast every type of cluster value are considered once. Splitting of Network Matrix (NetworkMatrix) is intended so that we can supervise the model with the trainset matrix to predict the testset matrix. It is variable to adjust the amount of train data and test data. It takes dataset as input and the parameter to divide data and then yields output as trainset and testset matrix that stores link between the users. trainset_nodes, testset_nodes contains the respective nodes.

Input: NetworkMatrix

[trainset,testset]= create traintestset(NetworkMatrix);

Return trainset, testset

Output: trainset, testset

Step2 Calculation of Degree and Max Degree- In this step calculation of degree of every node of the trainset_node is calculated. Degree is defined as the number of nodes that are linked or we say the number of neighbours. Max_Degree is the term that is assigned with the maximum value of all degrees in the trainset_nodes.

Step3 Proposed Step- In this step, complete working of the logic is described.

3.1 Calculation of number of cluster- It takes threshold value which is a value that can be calculated by taking the average of minimum degree and maximum degree; and degree of every node from step2 as input.

Number of clusters(k) = count /* count number of nodes

where {degree(node) > threshold value}

for each node € trainset_nodes (equation1)

3.2 Perform clustering using CBLP-

Algorithm1: CBLP

Input: trainset_nodes, trainset, k **Output:** Q[N,2], cluster[k,3]

1. N= size(trainset_nodes)

/* it computes the number of nodes in the trainset

2. k= Calculate the number of clusters using equation1

3. Random place k points on the space represented by the nodes of the trainset. These points are marked as initial cluster centres centre_value

4. for each (node € trainset_nodes)

Assign node to the group that has closest centre.

5. Label groups as cluster i

where i=1 to k

6. for i=1 to k

a.) calculate mean_value = mean of distances between the node and the centre of the cluster i; for every node € cluster i

b.) update the centre_value of cluster i = point at mean_value from all intracluster nodes

c.) cluster[i][1]= i

cluster[i][2] = centre _value[i][1]

cluster[i][3] = centre_value[i][2]

/* stores centre position of cluster

end for

7. Repeat step 4 to step 6 until the centres no longer move.

8. for j=1 to N

a.) Q[j][1]= trainset_nodes[j][1]

b.) Q[j][2]= i where i= label of cluster to which Q[j][1] belongs

end for

9. Return Q, cluster

Algorithm 1 is the CBLP method in which k, trainset are taken as input which is already discussed to produce Q, cluster. Q is the matrix that stores the indexes of all nodes that means the record of all nodes with their cluster value to whom the particular node belongs. Cluster contains the centre position of all the final updated clusters.

3.3 Calculation of user similarity

This is the step used to calculate the user similarity for all the non existent links using algorithm 2. If the simple clustering information of nodes in social network is used; it is insufficient to improve the performance of link prediction. To increase the accuracy and performance we presented a new user similarity measurement metric $S_{x,y}$ that includes the cluster information, global information of the network and the semantic similarity.

$$S_{x,y} = \theta * f1 + (1-\theta) * \frac{k_x * k_y}{(Max\_degree)^2} \quad \text{(equation 2)}$$

$$f1 = \alpha * \frac{com^{in}_{x,y}}{com_{x,y}} + (1-\alpha) * sem_{x,y} \quad \text{(equation3)}$$

here, θ,α are the flexible parameters that can be adjusted according to the value of factor that need to consider more. k_x defines the degree of the node x in the global network and k_y defines the degree of node y in the global network. Whereas Max_degree is the term that is assigned with the value of maximum degree of the network. $com^{in}_{x,y}$ is the number of common neighbours belonging to the same cluster with nodes x and y; this value become 0 when belongs to different clusters. $com_{x,y}$ is the number of common neighbours between x and y in the global network. And $sem_{x,y}$ is the score value calculated from the Semantic similarity matrix for node x and node y using algorithm 3 and maintained in sem[m][m] matrix. S is the matrix in which the user similarity scores are maintained. For nonexistent links; its value is predicted using algorithm 2 but for trainset data its value is 1. Whereas S1 is the matrix that stores the score 1 for the trainset data and else score 0.

Algorithm 3: Generation of Semantic similarity matrix sem[m][m]

Input: data set **output:** sem[m][m]

1. Read the data set and calculate m= size(trainset_nodes+testset_nodes)

2. Apply preprocessing of the data

3. for i=1 to m

for j=1 to m

a.) Apply wordnet similarity

b.) calculate score

c.) assign sem[i][j] = score

4. return sem[m][m]

Algorithm 2: Calculation of user similarity

Input: trainset, testset, trainset_nodes, testset_nodes, θ, α, cluster, Q, sem[m][m] **Output:**S[m][m], S1[m][m]

1. for ∀ (x and y)

if (x and y) ∈ trainset

S[x][y]=1

else

S[x][y]=0

2. for ∀ (x and y)

If(x and y) ∈ testset

S1[x][y]=1

else

S1[x][y]=0

3. Compute the degrees of all nodes in the trainset network

4. for ∀ (x and y) ∈ trainset_nodes

a.) Compute the user similarity $S_{x,y}$ using equation 2 and equation 3

b.) S[x][y] = $S_{x,y}$

5. return S[m][m], S1[m][m]

Step 4. Evaluation of AUC- This is the step use to evaluate the accuracy of link prediction with AUC value. This is the pre-existing algorithm to calculate the AUC[10].

Algorithm 4: Evaluation of AUC

Input: S, S1, testset **Output:** AUC

1.Num=0 , Morenum=0, Equalnum=0;

2. Diffset= S – trainset;

/* computes the different set between

3. for each link in testset

```
    for each link in Diffset
      if S1(i,j) in testset > S(i,j) in Diffset
          Morenum = Morenum + 1; Num = Num + 1;
      else if S1(i,j) in testset == S(i,j) in Diffset
          Equalnum = Equalnum + 1; Num = Num + 1;
     else
         Num = Num + 1;
      end
    end
end
```

4. AUC = (Morenum + 0.5* Equalnum)/ Num;
5. return AUC;

Step 5 Link Prediction- This is the step to predict the non existing links using the user similarity calculated in Algorithm 2 in Step 5. For a user x, from the user similarity; top users

are selected with whom user x has greater similarity value.

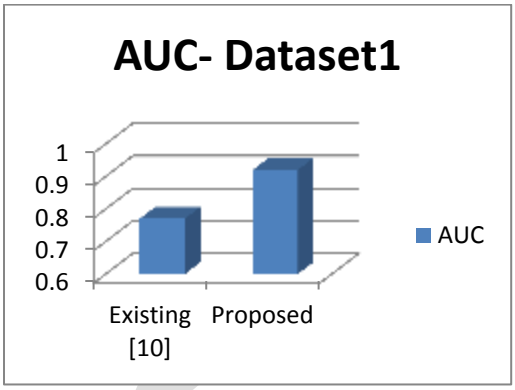
Step 6 Recommendation- For a user x, check the non- existing links and then considering its top similar users; it is recommended with the user that are not linked. It is believed that higher a similarity exist between two users; higher is the probability to link in future. With generating more accurate user similarity; it aims to recommend more accurate users.

5. Experiment/Performance Evaluation

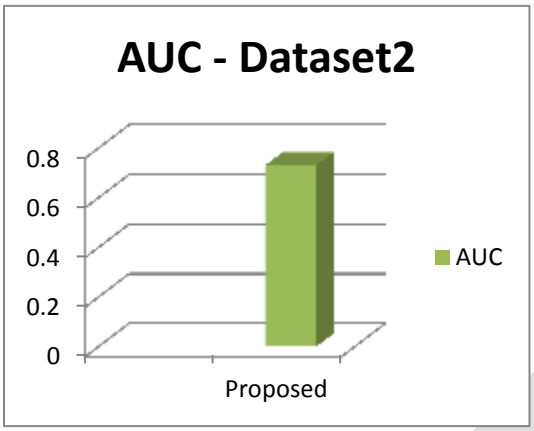
In this section, we firstly perform experiments on the real world datasets and then validate the performance of our proposed method.

5.1 Datasets

There are two datasets that are considered in our experiment. Firstly, is the karate dataset1 which is a social network data of interactions between members of a karate club by Wayne Zachary[ ]. This is the real world data set. Secondly, is the Custom built user defined dataset2 of a social network that includes a.) the links between the users and b.) the text messages, text comments that are shared by a particular user. In karate there are 34 users and 78 links. In custom built dataset there are 40 users and 114 links.



5.2 Results



In our experiments, we find AUC value for the best prediction performance on the above two mentioned datasets. Existing value is the AUC value for the existing clustering technique in existing work[10]. For dataset1, we are not considering factor semantic as it is a numeric data. And for dataset2 we considered semantic factor with the clustering technique. In existing work, AUC of dataset1 is 0.70 and AUC of dataset1 with proposed technique is 0.92 .

6. Conclusion and Future Work

This is concluded that we achieved our aim both the justifiability and high accuracy in link prediction. In order to achieve the goal we presented a CBLP algorithm which is based on clustering; considering both intracluster and topological structure information. With the efficient selection of number of cluster; we can improve the accuracy of link prediction in social networks. With clustering, semantic factor also enhance the AUC factor in text based datasets.

In the future work, we can examine the work on the more complex datasets from various domains and work with map reduce dimensionality and can improve the user similarity matrix by considering more unique factors.

REFERENCES:

[1] Pallavi Gupta, Sanjiv Sharma, "A Survey On Link Prediction Problem In Social Network", International Journal for Science And Research In Technology (IJSART) volume 1 Issue 1–JANUARY 2015.

[2] Zan Huang, "Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient" - LinkKDD 2006.

[3] D. L. Nowell and J. Kleinberg, "The link-prediction problem for social network," Journal of the American Society for information science and Technology, vol. 58, no. 7, pp. 1019-1031, 2007.

[4] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. , Science Direct,July,2003.

[5] Hossam S. Sharara Walaa Eldin M. Moustafa, Link Prediction, www.cs.umd.edu/class/spring2008.

[6] Svitlana Volkova, "Link Prediction In Social Network".

[7] Mohammad Al Hasan, Mohammed J. Zaki, "Link Prediction In social networks", Springer,2011.

[8] SAP HANA Predictive Analysis Library (PAL), 2015.

[9] Linyuan Lu, Tao Zhou, "Link Prediction In Complex Networks", Elsevier 2010.

[10] Fenhua Li, Jing He, Guangyan Huang,"A Clustering-based Link Prediction Method in Social Networks", International Conference on Computational Science, Procedia Computer Science,2014.

[11] Jorge carlos valverde-Rebaza, Link Prediction in complex networks based on cluster information, Elsevier.

[12] Jungeun Kim, Link Prediction Based on Generaliz ed Cluster I nformation.

[14]Lü L.,Zhou T. Link prediction in complex networks: a survey. Physica A: Statistical Mechanics and its Applications, 390:1150-1170,2011.

[15]Huang Z.,Li X.,Chen H. Link prediction approach to collaborative filtering. in Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 141-142: ACM, 2005.
[16]Newman M. E. Clustering and preferential attachment in growing networks. Physical Review E, 64, 2001