Research Journal of Computer and Information Technology Sciences ______________________ISSN 2320 – 6527
Vol. **3(2),** 6-10, May (**2015**) *Res. J. Computer and IT Sci.*

Combining Jaccard Coefficient with Fuzzy Soft Set for Predicting Links in Social Media

Yachana Bhawsar, G.S. Thakur and R.S. Thakur
Department of Computer Application, Maulana Ajad National Institute of Technology, Bhopal, MP, INDIA

Available online at: www.isca.in
Received 2nd May 2015, revised 15th May 2015, accepted 20th May 2015

Abstract

Link prediction in social networks is not a new research area but relations on social sites are growing very fast. So many link prediction techniques have been developed. Building a relation on social media is not a certain event. Fuzzy soft set is a suitable approach to handle such uncertainties. In this paper fuzzy soft set based link prediction model is proposed. In this model various features of social networks are used. We combine existing similarity score named jaccard coefficient with fuzzy soft set to predict the correct link. The comparative analysis has been done with the existing methods. The efficiency of the proposed method is better than the existing methods of link prediction like common neighbor, jaccard, Sorenson etc.

Keywords: Social network, link prediction, fuzzy soft set.

Introduction

Social networks become very popular between users. All the websites wants to attract people to join their network and they have its own kind of services for this purpose. Regular growth and generation of new connection is basic requirement of the network. Here growth means joining of new user in the network and generation means developing of new links between users of same network. Classification and association rule mining are widely used data mining techniques[1]. These techniques play important role in link prediction. This paper takes link generation as the main issue because most of the social sites like Facebook, Linkedin, Google Circle etc. always suggest their user for new links. If more number of friends are suggested by any social networking sites then users will not move to other similar network where less friends are available. For making strong prediction of friends, prediction algorithms use various feature of social network. Expert systems and artificial intelligence are used for making such kind of decisions[2]. Researchers investigated the relations of human resource development and organizational development[3]. This is the basis for User-User relations used in this paper. In social networks users have to make smart decisions and user's behavior is very dynamic and challenging. This built up a strong competition for social network to sustain their position[4]. Some researchers found that the paradigm is shifting from product to customer orientation[5].

So the recourses of interaction among people are online social networks. Social network can be visualized as a graph g. As we know there are some vertices and edges in the graph. So the vertex corresponds to the person in that particular network. Usually vertices are represented by a set V. The relation between two persons can be represented by the edge between those persons. So relations are represented by the set of edges E. Predicting links in such a dynamic social network is challenging task. Link prediction can be understood by figure 1 where each vertex represents user while each edge represents their friendship. So if two vertex have connection then they are friends otherwise there may be chance that they will become friends in future. In figure 2 new links are predicted and represented by red edge i.e. between user A and user B.

Researchers found that there are three types of link prediction models exist. Some model extracts a set of features to train a binary classification model, and these are traditional or non-Bayesian models. Some models model the joint probability among the entities of a network, and these are probabilistic or Bayesian graphical models. Some models compute the similarity between the nodes of the network and based on rank-reduced similarity concept, and these are liner algebraic models. The problem of link prediction is solved by defining proximity-based measures on the nodes in the underlying network[6-8], usefulness of different topological features[9], statistical relational models[10-14], context of the classification problem[15,16,17]. These existing methods have some limitations. Methods those are based on topological features are not valid for new nodes. Classifier approaches fails on reliability of training data. Applying fuzzy soft set for link prediction is a new method.

Related work

With the increase of social network on internet, various researchers have elaborate different fields. Out of those, link prediction is playing an important role. As the network is very vast and dynamic, single technique is not sufficient for prediction. The earliest link prediction model for social network is proposed by Liben-Nowell and Kleinberg[6]. Basically link prediction model extracts the similarity matrix between the pair of vertices. Graph based similarity metrics are used to calculate similarity index. After calculating similarity index, ranking of the similarity scores is done to predict the link between two vertices. Various topological features like graph shortest distance, common neighbours, preferential attachment, Adamic-Adar, Jaccard, SimRank, hitting time, rooted Pagerank and Katz are examined by them. Various graph based similarity matrices for link prediction are examined by them. Later, Hasan et. al.[17] extended this work in two ways. According to them we can improve prediction result by using external data outside the scope of graph topology. And as a second way we can used various similarity metrics as features. Here link prediction problem can be treated as binary classification problem. Since then, the supervised classification approach has been popular in various other works in link prediction.

The available methods such as decision tree induction, naive Bayes[18], support vector machine, logistic regression, etc. are examples of supervised learning techniques. Doppa et. al. proposed a learning algorithm for link prediction based on chance constraints[19]. Supervised learning is based on feature construction. After construction of features collective classification of learned model is the next step of the model. Feature vector is referred to as the computed features for a particular node pair. Feature vector is correlated with the future possible link between that node pair. We have the set of computed feature vectors for training data. And then with the help of these feature vectors we train the learning model. Then the model is used to predict the future links[20].

Most of the approaches like common neighbour, jaccard, Sorenson uses single feature for predicting link. User-User relation from social network has sufficient information for prediction. The prediction accuracy of existing methods is very low due to the information available for prediction is in the form of features and became useless. Multiple features are used by Markov Model in[21] for increasing prediction accuracy but values obtain from different features are not normalized. One more problem with Markov modal is sequence pattern followed by the user is used for prediction, which is highly vulnerable as user has different priority for different user. In this work we applied user-user dataset and different features. For generating fuzzy values we used jaccard coefficient and effective normalization method. This will improve prediction accuracy of the work.

Proposed Work

In this work User-User relation is used for predicting links between users. A new combination of features is generated for link prediction, which is based on these relations. To understand user-user dataset two users A and B are taken for example. If A and B are two users in any social networking site, and then they may have relation in terms of like, comment, image share, video share, message, comment share, friend request, same group, common friends, video chat, text chat etc. These may be different activities between two users in any social network. One user can perform these activities any number of times. So we count the number of times these activities performed between two users like user A and user B. To understand User-User relation we are taking the example network represented as graph in figure-1. There are seven nodes in the given social network i.e. seven users in the network. We are taking six features of social network, these feature may be like, comment, message, post, share, friend request and so on. Out of many features we are taking six appropriate features between any two users. Table 1 is showing the tabular representation of User-User relation between user A and remaining six users of example network. If user A has performed any activity to any other user than we have to count which activity performed how many number of times and enter that data to the appropriate place in the table. So table 1 showing the data for node A in our example network. If there is zero in table that means that particular activity is not exist between those users. We can see in table-1 that lots of zeros are there. But if we talk about user A and user C than there are count for each and every feature. This happens because user A and user C are connected n figure-1. There are some counts for some features between user A and User B, it shows that there are chances to be a link between these two users. Other features are having zero counts with respect to A and other user i.e. D,E,F, and G that means there are lesser chance to be a link between that user-user.

Table-1
User-User relations with feature counts

User	User	Feature-1	Feature -2	Feature-3	Feature -4	Feature-5	Feature-6
A	B	0	2	0	5	0	0
A	C	10	2	4	6	3	2
A	D	0	2	0	3	0	0
A	E	1	0	4	0	0	2
A	F	2	1	1	0	0	0
A	G	0	0	0	2	2	1

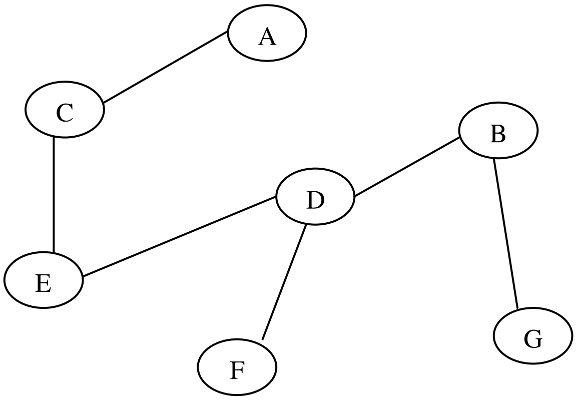

Figure-1
Graphical Representation of Social network

So in this way all the users have feature counts for other remaining users. The frame work of the proposed work can be understood by figure 3. In the next section we briefly explained each step of the proposed model.

Pre-processing: As explained above the user-user dataset contains number of feature between users, so conversion of this dataset as per working environment should be done. In this step dataset is arranged into matrix in which first two columns represent user-id, while rest of the columns represent the feature count values. If zero is present in the column then it shows that the particular feature is not used by the specified user ids.

Jaccard Coefficient: In order to find strong relationship values between users, jaccard coefficient is used in the proposed work. Jaccard is a similarity index user in many link prediction algorithms. Here we are using jaccard coefficient to calculate a measure of friendship. The concept is same as the jaccard coefficient but the application of it is different. If A and B are two users then A∩B shows the number of times any feature used by A to B or B to A, while AUB means total number of times any feature activity done by A and B to other friends. So this will give a single value for a particular feature between A and B. In this fashion all feature value can be obtained by jaccard coefficient using equation-1. For example if we talk about 'like' feature. Then let A∩B has 5, while AUB has value 50 then Jaccard Coefficient value for 'like' feature is 0.1. In e quation 1 m represents number of that particular feature.

$$J_{x,y,m} = \frac{X_m \cap Y_m}{X_m \cup Y_m}$$ equation (1)

Normalization: Different features have different priority. Normalization is required to put the entire feature at same scale. A weight vector which contains values for the entire feature is used for normalization of the various values. This makes all features at same level. So if J is jaccard coefficient matrix then normalization is done by equation 2. In above equation Wm is weight matrix value for mth feature.

$$N_{x,y,m} = J_{x,y,m} \times W_m$$ equation (2)

Decision factor: The decision factor[22] obtained after the normalization has separate values for each feature, and then it is required to make single value for the decision. So one comparison matrix is developed, where each value of the features are pass through the equation 3. Here X, Y represent users while m represents number of features and J represents jaccard coefficients.

$$C_{x,y} = \sum_{k=1}^{m} (N_{x,y,k} - N_{y,x,k})$$ equation (3)

Now each user has single value based on the user-user relation. This matrix is utilized in the decision function of equation 4, where summation of each is done corresponding to the users.

$$R_x = \sum_{y=1}^{n} C_{x,y}$$ equation (4)

Proposed Algorithm

As mentioned above the proposed algorithm will use User-User data and will give the predicted links as the output. The work is combining the jaccard coefficient with fuzzy soft set. In this the preprocessing of data is done in step 1 then we worked as the proposed framework suggests i.e. we calculate jaccard coefficient for each use- user relation. For each feature between two users we have to normalize the calculated jacaard coefficient. In step 3 and 4 we do the same. Then for each friend and for all features we have to construct comparison matrix. After constructing the comparison matrix we calculate the decision factor.

```
Input: UUD
Output: Link_Prediction
UUD ←Pre-Processing(UUD)
JC←Jaccard_Coefficient( UUD) // equation 1
Loop 1: m // m represent number of features
    N[m] ← Normalization(JC[m]) // equation 2
End Loop
Loop 1: n // n represent number of friend
    Loop 1: m
        CC[n] ← Comparision_Matrix (N[n, m]) // equation 3
    End Loop
End Loop
Loop 1:n
    DF[n] ← Decision_Function (CC[n]) // equation 4
End Loop
Link_Prediction←Sort(DF)
```

Experimental Results

In this section experimental results of proposed work are shown. The implementation of algorithm and utility measures are done on MATLAB tool. The tests were performed on a 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional. Comparison of proposed work is done on various evaluation parameters such as Precision, Recall and F-score and area under the curve. Measuring the classification accuracy involves some measures. Here the dataset is divided into training and testing data. The model is built on training data and it is tested on testing data. The accuracy of the proposed model is explained by true positive (tp), true negative (tn), false negative (fn), and false positive (fp). These quantities are explained in the table-2. This is also called confusion matrix. We have to count following quantities to show the prediction accuracy of the model. These quantities are described as follows:
Precision = tp/(tp+fp), Recall= tp/(tp+tn), F-measure= 2*Precision*Recall/(Precision+Recall)

Table-2
Confusion Matrix

	Friend ((Predicted)	**Not friend (Predicted)**
Friend (Actual)	True positive (tp)	False negative (fn)
Not friend (Actual)	False positive (fp)	True negative (tn)

Improving Recall allows the longer friend lists and it reduces precision. There are some curves that compare precision to recall and called precision-recall curves. These are some curves comparing true positive rate to false positive rates and called receiver operating characteristic or ROC curves. F-measures and Area Under the ROC curve (AUC) summarize the precision recall of ROC curve.

For experiment 200 users and 14 events is taken from dataset which contains total 24000 transactions. For comparison 10000 transactions are used for training of different methods and testing is done on rest. Proposed work is compared with four existing methods Markov Modal, Common neighbour, Jaccard, Sorenson.

From table 3 it is obtained that the values of all parameters of proposed work for link prediction is much better as compare to other existing algorithm. Value shows that the result of proposed work is two times accurate than the previous work i.e. by Markov modal and many times as compare to other existing methods. It is observed from the figure-2, that AUC of fuzzy soft set is 0.044 which is much high as compare to all methods.

Table-3
Evaluation parameters for data set size 10000

Method	**Precision**	**Recall**	**F-measure**
4th order Markov	0.1967	0.0737	0.1073
Common Neighbor	0.0305	0.0299	0.0302
Jaccard	0.0284	0.0278	0.0281
Sorenson	0.0101	0.0099	0.1000
Fuzzy soft set	0.3831	0.2303	0.2877

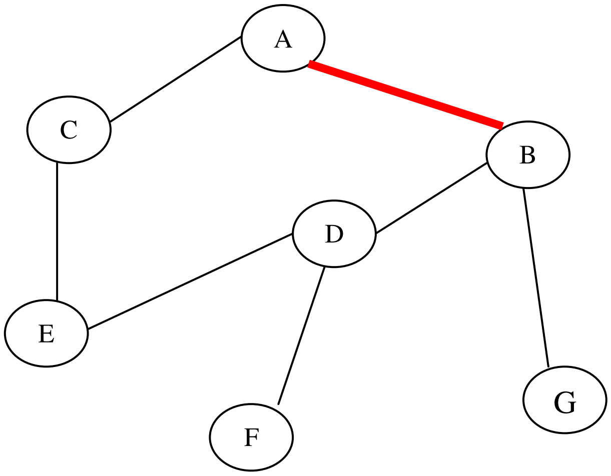

Figure-2
Graphical Representation of Social Network with new link between user A and user B

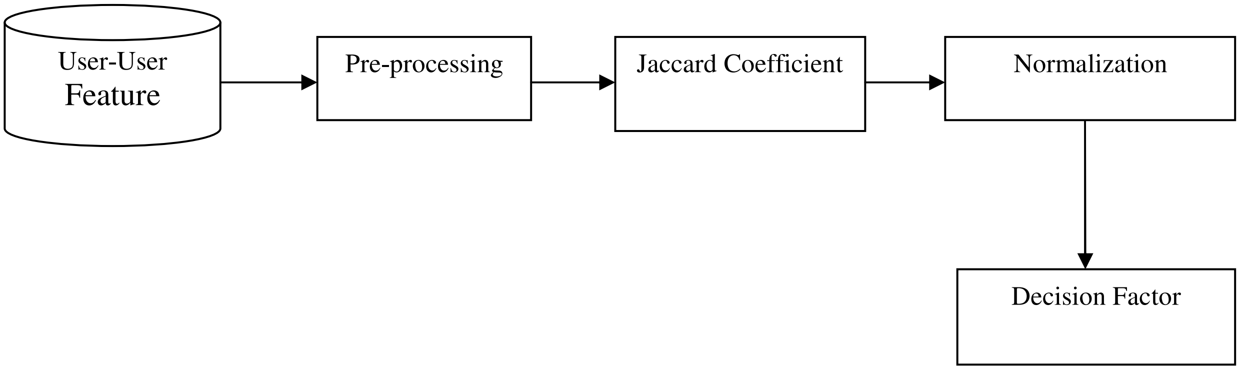

Figure-3
Block diagram of Proposed Work

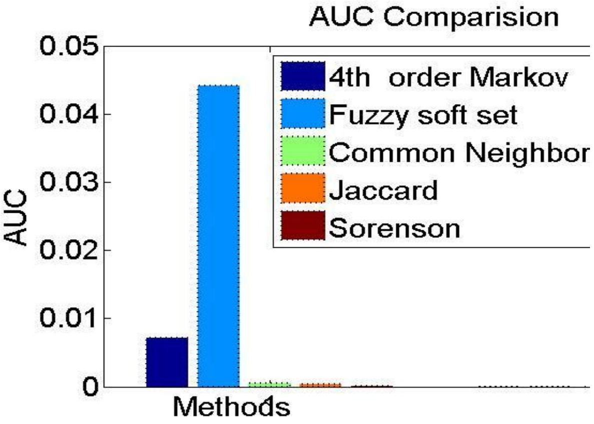

Figure-4
Area under curve (AUC) values comparison of different methods

Conclusion

Link prediction for user is an important tool for social network. Keeping this goal in mind proposed work uses a combination of jaccard index coefficient and fuzzy set for link prediction. Different features from the social network is utilized for prediction, which get normalized to common platform depends on their priority; it was not done by any other previous approaches. Results obtain after experiment is compared with various approaches such as Markov, common neighbour, etc. It has been found that fuzzy soft set is much better than all other approaches on different evaluation parameter. In future a combination of two or more social network is required, as people on different network have more chance to get together in coming time.

References

1. Varsha Namdeo, R.S. Thakur and G.S. Thakur, Self Organized Map Network for Classification of Multilevel Data, *Research Journal of Computer and Information Technology Sciences,* **3(1),** 1-4 **(2015)**
2. Reza Khodaie Mahmoodi, Sedigheh Sarabi Nejad and Mehdi Ershadi sis, Expert Systems and Artificial Intelligence Capabilities Empower Strategic Decisions: A Case study, *Research Journal of Recent Sciences,* **3(1),** 116-121 **(2014)**
3. Droudi Homa and Dindar Farkoosh Firouz, An Investigation on the Relation between Human Resources Management and Organizational Developments, *Research Journal of Recent Sciences,* **2(2),** 50-53 **(2013)**
4. Raorane A.A., Kulkarni R.V. and Jitkar B.D., Association Rule – Extracting Knowledge Using Market Basket Analysis, *Res. J. Recent Sci.,* **1(2),** 19-27 **(2012)**
5. Masume Sadat Abtahi, Mohammad Bayat and Ali Abolghasemi., Model Successful Implementation of Customer Relationship Management (Case Study: Ghavamin Bank), *Research Journal of Recent Sciences,* **4(2),** 130-138 **(2015)**
6. D. Liben-Nowell and J. Kleinberg, The link prediction problem for social networks, LinkKDD, **(2004)**
7. Lada A. Adamic and Eytan Adar, Friends and neighbors on the web, *Social Networks,* **25(3),** 211-230, **(2003)**
8. M. E. J. Newman, Clustering and preferential attachment in growing networks, *Physical review Letters,* **(2001)**
9. O. Nasraoui and R. Krishnapuram, One step evolutionary mining of context sensitive associations and Web navigation patterns, in Proc. SIAM Int. Conf. Data Mining, Arlinton, VA, 531-547, **(2002)**
10. A. Popescul, L. Ungar, S. Lawrence and D. Pennock, Statistical relational learning for document mining, ICDM, **(2003)**
11. B. Taskar, P. Abbeel, and D. Kollerk, Discriminative probabilistic models for relational data, UAI, **(2002)**
12. L. Getoor, N. Friedman, D. Koller, and B. Taskar, Learning probabilistic models of relational structure, ICML, **(2001)**
13. M. Bilgic, G. Namata and L. Getoor, Combining collective classification and link prediction, Workshop on Mining Graphs and Complex Structures, ICDM, **(2007)**
14. O. Hassanzadeh and et al, A framework for semantic link discovery over relational data, CIKM, **(2009)**
15. C. Wang, V. Satuluri and S. Parthasarathy, Local probabilistic models for link prediction, ICDM*,* **(2007)**
16. H. Kashima and N. Abe, A parameterized probabilistic model of evolution for supervised link prediction, ICDM, **(2006)**
17. M. Al-Hassan, V. Chaoji, S. Salem, and M.J. Zaki, Link prediction using supervised learning, workshop on Link Analysis, Counterterrorism and Security, SDM, **(2005)**
18. Jiang, L., Zhang, H., Cai, Z. Discriminatively Improving Naive Bayes by Evolutionary Feature Selection., *Romanian Journal of Information Science and Technology,* 163–174, **(2006)**
19. Janardhan Rao Doppa, Jun Yu, Prasad Tadepalli, and Lise Getoor. Learning algorithms for link prediction based on chance constraints. In Proceedings of European Conference Machine Learning and Knowledge Discovery in Databases, 344–360, **(2010)**
20. Milen Pavlov and Ryutaro Ichise., Finding experts by link prediction in coauthorship networks., In Proceedings of 2nd International ISWC ASWC Workshop on Finding Experts on the Web with Semantics, 42–55, **(2007)**
21. T. Joachims, D. Freitag, and T. Mitchell, WebWatcher: A tour guide for the World Wide Web, in Proc. IJCAI, 531-547, **(1997)**
22. P.K. Maji, A.R. Roy, R. Biswas, An application of soft sets in a decision making problem, Comput. Math. Appl., 1077-1083, 44 **(2002)**