Friend Recommendation System for Online Social Networks using Cohesion Based Approach

Sanjeev Dhawan

Faculty of Computer Science and Engineering, Department of Computer Science and Engineering, University Institute of Engineering and Technology (UIET), Kurukshetra University, Kurukshetra (KUK), Haryana, India

Kulvinder Singh

Faculty of Computer Science and Engineering, Department of Computer Science and Engineering, University Institute of Engineering and Technology (UIET), Kurukshetra University, Kurukshetra (KUK), Haryana, India

Ashu Jain

M.Tech (Computer Engineering) Student, University Institute of Engineering and Technology (UIET), Kurukshetra University, Kurukshetra (KUK), Haryana, India

Abstract – **Social Networks (SNs) are mostly used by people for communicating or sharing their emotions with each other. Recommending a friend to join new group or adding a new friend is a typical task. Every day various techniques are developed such as match maker or content based recommendation to recommend a friend. Each technique suffers from its pros and cons. This paper is divided into six sections. Section 1 and 2 discuss the friend recommendation and various existing techniques of the recommendation. Section 3 provides a brief overview of the work done so far in the field of friend recommendation and their shortcomings. Section 4 discusses the proposed mechanism and the results are described in the section5. Finally the section 6 discusses the conclusion drawn from the whole work. In this paper an attempt has been made to propose a novel cohesion based friend recommendation system to recommend new friends based on some existing parameters such as mutual friends, common workspace and age of person. This work has been implemented in Eclipse and analyzed using Weka.**

Index Terms – Social Networks, Weka, Cohesion.

1. INTRODUCTION

Online social network is an online service that simulates the human social interactions and relations of real life. It allows users within the social network to communicate with other users interact with them and add them to their friends' lists. With the rise in popularity of online social networks, many other types of sites began to include social networking features. Examples include multimedia content sharing sites (Flickr, YouTube, and Zoomr), blogging sites (Live-Journal and BlogSpot), professional networking sites (LinkedIn and Ryze), and news aggregation sites (Digg, Reddit, and delicious). All of these sites have different goals but employ the common strategy of exploiting the social network to improve their sites. The list above is not meant to be exhaustive, as new sites are being created regularly [1].

2. FRIEND RECOMMENDATION SYSTEM

In recent times, people of almost all age group are connected to each other through the means of social networks. This connection of people is not limited to the boundaries of known and unknown. Apart from connecting to the known world, people tend to connect to the other people they may not know in the real world. Thus to ease this process of connecting to the persons all over the world, Friend Recommendations systems are developed which recommend friends on the social media using attributes likes common interests or same profession or common likes. Services on online social networks like 'Friend Suggestions' or 'Similar People' or 'People you may know' are using this Friend Recommendation System for suggesting friends to the users. Friend Recommendation Systems make use of several filters to separate the not-recommended users from recommended ones.

A. Existing Techniques

1) Match Maker: It is an automated collaborative filtering based mechanism which recommends friends based on similarity with the TV actors. In figure 4.1, A and B are TV actors who are friends to each other. User1 resembles character A. Also, User2 is similar to character B. So the system recommends User1 and user2 to become friends, if they are not already friends. This technique was applied on Facebook. It is just a personality matching system and ignores proximity matching [2].

2) Content-based Recommendation: In this type of recommendation system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes. It tries to recommend items that are similar to those that a user liked in the

past. Some movie recommendation algorithms using content-based filtering are: Rotten Tomatoes, Internet Movie Database etc. In content-based recommendation system, users' past likes and interests are taken into account [3].

3) Geographical Recommendation: In this type of recommendation systems, users that live geographically close to each other are recommended while those far apart have very less chance of recommendation to each other. Geographical Recommendation system does not take into account users' likes and dislikes [4].

3. BACKGROUND WORK

The graph terminology for social networks was used by Wang *et al.* 2013 [5] where the problem was defined as a graph with the users and the locations as the graph nodes, the friendship relation as the user-user edges, and the user-location relation as the user-location check-in edges. The friendship based similarity was computed by starting from the target user and by ranking all the users (that formed the user-user link). Similarly, Dhawan and Goel 2014 [6] examined the privacy issues on usage of social networking sites. They provide analysis on detection of fake users by checking their valid Id proof that whether they are legitimate users or not. Similarly Dhawan *et al.* 2015 [7] designed a high rating preference system based on item to item collaborative filtering technique. Yang *et al.* 2014 [8] presented a survey on collaborative filtering based social recommender systems and concluded that a social recommender system improves on the recommendation accuracy of the traditional systems by taking social interests and social trusts between users in a social network. In addition, Wu *et al.* 2015 [9] proposed a friend recommendation algorithm based on a user similarity graph in order to locate candidate friends who share interests in social tagging systems such as Delicious and Last.fm. They used the latent Dirichlet allocation (LDA) and a multi-view users' similarity measure between two users based on their topics of interest.

Many of these algorithms suffer from the time complexity issues and also the past history of the users was not considered before recommending the friends. Hence to tackle these issues, cohesion based approach is proposed in this paper.

4. PROPOSED WORK

Cohesion is defined as the sum of all factors that attract some person to join a particular group or an organization. Cohesion can be thought of as the sum of two factors-density and connectedness. Density defines the modularity of the network or graph and modularity measures the strength of partitions of network into the modules. Connectedness defines how well connected the network is [10]. In the proposed work, cohesion of two or more attributes is used to make the decision of recommendation. Fig. 1 describes the working of the proposed

work. This work is proposed for a user who travels to a different country for study and recommends the local residents as friends to the user.
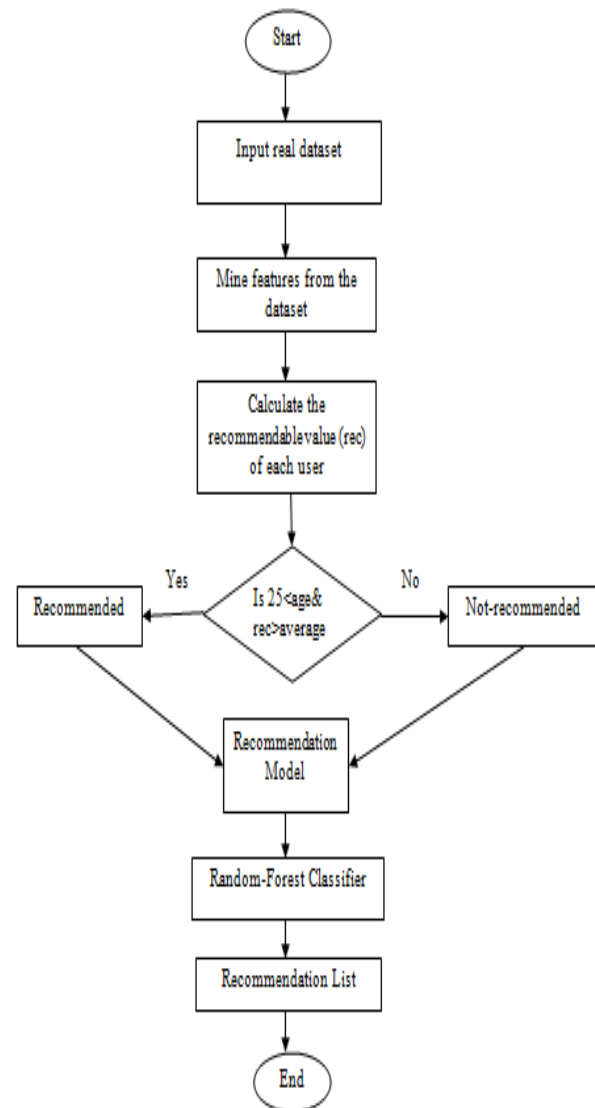

Figure 1: Flow chart of proposed system

Initially the real dataset from twitter was used to mine features. Using these features, a new parameter to recommend the friends, Recommendable Value (rec) is calculated using the following formula:

$$rec = \sum_{i=0}^{n} w_i P_i \qquad (1.1)$$

Here n defines the total number of users in the dataset, w_i defines the weight of the ith parameter and P_i defines the Parameter value of the parameter i. Hypothetical weights are assigned to the parameters which shows their contribution in decision making. A weight of 0.5 is assigned to the Trust Value

and 0.2 is assigned to Mutual Friends. Based on these values, value of rec is calculated and if the value of rec>average value and age of the person is greater than 25, the person is recommended to the user. Finally a recommendation model is developed using these parameters and fed to the Random Forest Classifier to evaluate the performance of the model using metrics like TP rate, FP rate, and Precision, Recall, F-measure and ROC curve.

A. Random Forest Classifier

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of over fitting to their training set. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, …, x_n$ with responses $Y = y_1, …, y_n$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples:

1) For $b = 1, …, B$:

2) Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

3) Train a decision or regression tree f_b on X_b, Y_b.

4) After training, predictions for unseen samples x' can be made by averaging the prediction from all the individual regression trees on x' or by taking the majority vote in the case of decision trees [11].

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x') \qquad (1.2)$$

5. RESULTS AND ANALYSIS

The proposed work is carried out in the Eclipse software configured using Weka software tool.

A. Weka

Weka (Waikato Environment for Knowledge Analysis) is an open source software tool written in Java. It was developed at University of Waikato, New Zealand to work with .arff (Attribute Relation File Format) files. It provides various mechanisms for data processing like preprocessing, classification, clustering etc [12].

B. Eclipse

Eclipse is an open source software tool developed from the Prolog programming language. It is generally used in the areas of resource allocation, scheduling, timetabling, planning, transport etc. It has a lot of pre installed libraries, interfaces to third party solvers, a high-level modeling and control language and an integrated development environment [13].

C. Performance metrics used

To measure how well the designed system is performing, a social network metric called F-measure or F1 score is used, which can be calculated using two other metrics called precision and recall.

$$precision = \frac{TP}{TP+FP} \qquad (1.3)$$

and

$$recall = \frac{TP}{TP+FN} \qquad (1.4)$$

Thus, F-measure can be defined as:

$$F = 2 * \frac{precision*recall}{precision+recall} \qquad (1.5)$$

Where

1) True Positives (TP): Terms that were correctly identified as entities.

2) False Positives (FP): Terms that were identified as entities but should not have been.

3) True Negative (TN): Terms that were not identified as entities and should not have been.

4) False Negative (FN): Terms that were not identified as entities but should have been.

Confusion Matrix: It is defined as a table M that is used to describe the performance of a classification model on a set of data values for which true values are known.

TP	FP
FN	TN

ROC Curve: ROC is a receiver operating statistic curve that is used to illustrate the diagnostic ability of a binary classifier as the threshold values are varied. It determines the accuracy of the system. A value of 1 indicates the system is working with complete accuracy. It is plotted by taking False Positive Rate along x-axis and True Positive Rate along Y-axis [14]. The ROC Area for the proposed system is 0.985 as is shown in figure 2.

Table 1 describes the various performance metrics derived from the proposed work. The results described here depict various performance metrics for both the classes Recommended (Yes) and not recommended (No). As shown in the table 1 above TP rate is close to 1 and FP rate is close to zero for both the classes of recommendation. This shows that the performance of the proposed mechanism is very fair and also the value of F-measure is also close to 1.

Table 2 shows the confusion matrix for the proposed mechanism. Number of True Positives entities are 100 which shows that the proposed mechanism recommends only 100 users out of the users. Thus the proposed mechanism has optimized the performance of the friend recommendation framework up to some extent.

TABLE I

PERFORMANCE METRICS FOR PROPOSED MECHANISM

TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Classes
0.917	0.019	0.926	0.917	0.922	0.985	Yes
0.981	0.083	0.978	0.981	0.98	0.985	No

TABLE II

CONFUSION MATRIX FOR THE PROPOSED WORK
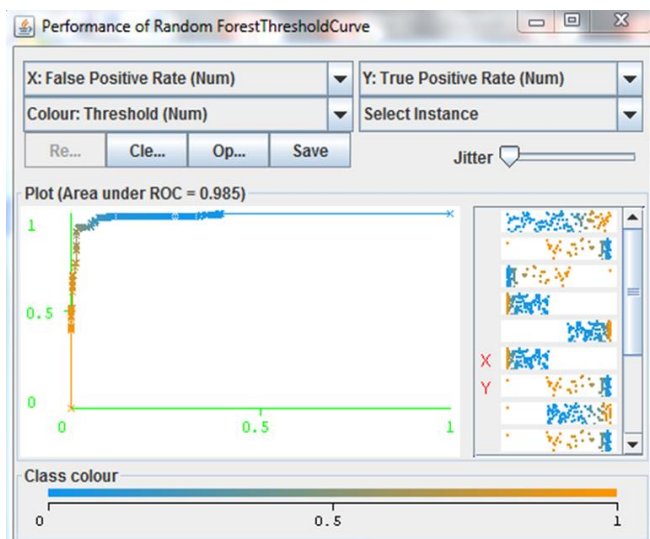
Yes	No
100	9
8	407

Figure 2: ROC Curve for the Proposed Mechanism

6. CONCLUSION

In today's world of e-commerce and e-business, people devote most of their time to the online services. Thus social networks also gained importance because of their desire to connect to a large number of people around the world. Friend recommendation on these social networking sites is a very difficult problem because usually people tend to connect to the people who might be similar in their tastes. Also the recommended user should be trustable because the user doesn't want to connect to some fake user. In order to overcome this kind of problem in this paper, a friend recommendation system called cohesion based recommendation system is recommended in which a new friend is recommended based on some trust value calculated on the basis of few existing parameters. Results shows that proposed system recommends more friends in an optimal way. In future, other classifiers can be used and the results obtained from each of them can be compared to see which classifier produces the better result. Additionally, daily lives of users can be modeled as life documents for this system that can track social activities of users repeatedly after specific interval of time to make friend suggestion more efficient.

REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of Computer-Mediated Communication 13(1):210-230, 2008.

[2] Li Bian, Henry Holtzman, Tuan Huynh and Marie-Jose Montpetit, "MatchMaker: A Friend Recommendation System through TV Character Matching", IEEE Conference on Consumer Communications and Networking, Las Vegas, USA, 2012.

[3] Michael J.Pazzani and Daniel Billsus, "Content-Based Recommendation System", Volume 4321 of the series Lecture Notes in Computer Science, pp 325-341.

[4] Jia-Dong Zhang, Chi-Yin Chow and Yanhua Li, "iGeoRec: A Personalized and Efficient Geographical Location Recommendation Framework", DOI 10.1109/TSC.2014.2328341, IEEE Transactions on Services Computing, 2015.

[5] Hao Wang, Manolis Terrovitis and Nikos Mamoulis, "Location recommendation in location-based social networks using user check-in data," in 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, Florida, November, 2013.

[6] Sanjeev Dhawan and Shivi Goel, "Analysis of Pattern of Information Revelation and Site Use Behavior in Social Networking Sites," International Journal of Computer Applications Technology and Research, vol. 3, no. 1, pp. 42-44, 2014.

[7] Sanjeev Dhawan, Kulwinder Singh and Jyoti, "High Rating Recent Preferences Based Recommendation System," in 4th International Conference on Eco-friendly Computing and Communication Systems, Kurukshetra, India, 2015.

[8] Xiwang Yang, Yang Guo, Yong Liu and Herald Steck, "A survey of collaborative filtering based social recommender systems," Elsevier, Computer Communications, vol. 41, pp. 1-10, 2014.

[9] Wu BX., Xiao J. and Chen JM, "Friend Recommendation by User Similarity Graph Based on Interest in Social Tagging Systems," in Advanced Intelligent Computing Theories and Applications, Liverpool, United Kingdom, 2015.

[10] Md. Nafiz Hamid, Md. Abu Naser, Md. Kamrul Hasan and Hasan Mahmud, "A cohesion-based Friend Recommendation System," Springer-Verlag Wien 2014.

[11] https://en.wikipedia.org/wiki/Random_forest accessed on 20th May, 2017.

[12] Ekta and Sanjeev Dhawan, "Classification of Data Mining and Analysis for Predicting Diabetes Subtypes using WEKA", International Journal of Scientific & Engineering Research, Volume 7, Issue 12, December-2016, pp: 100-103.

[13] https://en.wikipedia.org/wiki/Eclipse_(software) accessed on 25th May, 2017.

[14] http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ accessed on 30th May, 2017.