Research Journal of Applied Sciences Engineering and Technology 4(5): 500-509, 2012
ISSN: 2040-7467

Submitted: December 06, 2011 Accepted: January 10, 2011 Published: March 01, 2012

Hybrid Recommender System for Joining Virtual Communities

[1]Leila Esmaeili, [2]Behrouz Minaei-Bidgoli, [3]Hamid Alinejad-Rokny and [4]Mahdi Nasiri
[1]School of Computer Engineering, University of Qom, Qom, Iran
[2,4]School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[3]Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, Tehran, Iran

Abstract: The variety of social networks and virtual communities has created problematic for users of different ages and preferences; in addition, since the true nature of groups is not clearly outlined, users are uncertain about joining various virtual groups and usually face the trouble of joining the undesired ones. As a solution, in this study, we introduced the hybrid community recommender system which offers customized recommendations based on user preferences. Although techniques such as content based filtering and collaborative filtering methods are available, these techniques are not enough efficient and in some cases make problems and bring limitations to users. Our method is based on a combination of content based filtering and collaborative filtering methods. It is created by selecting related features of users based on supervised entropy as well as using association rules and classification method. Supposing users in each community or group share similar characteristics, by hierarchical clustering, heterogeneous members are identified and removed. Unlike other methods, this is also applicable for users who have just joined the social network where they do not have any connections or group memberships. In such situations, this method could still offer recommendations.

Key words: Collaborative filtering, content based filtering, entropy, recommender system, social network

INTRODUCTION

A social network is a social structure made of nodes that can be linked together. Links of financial interactions, friendship relationships, relations based on web and entertainment are examples of such links. In other words, social networks are areas in the cyber space where each user profile represents a user; this virtual world creates an opportunity for them to be linked with other people with whom they have common interests (Adamic *et al*., 2003).

People join social networks and along with many other useful and pleasing activities, make friends, join different social communities, build new relationships and play games. As human beings, users have different thoughts, preferences and interests; but those with the same ones join the same communities in the social network. These virtual communities could be created around many of fundamental ideas and obsessions of people like commerce, politics, society, religion and many other topics. Regardless of their types; however, there virtual communities and groups are increasing and many new ones are always being created. Due to their variety always there is a chance of joining wrong groups; and therefore, users need to be assisted in finding the right groups by receiving recommendations from the social network.

Recommender systems are nowadays an essential web application. By means of personalization in web and by using recommender systems, it is possible to recommend users, items, web pages, persons and groups that are based on their characteristics and needs. Among the recommendations made by the recommender system, the user is always free to choose. Information required in such systems is usually acquired from web usage analysis, web content, web structure and user profiles.

Unlike a movie recommender system or other similar systems in which the recommended item has features such as director, actors, genre and etc., here, a group contains members who have joined the group due to their common goal or interest; therefore, what is recommended in this system is different from recommendations made by other recommender systems, a group is not an object!. Regardless of its members, a group is the same as the features of its members; and we could identify a group by its members' features. We believe that users with the same features and interests join the same groups. Therefore, based on this hypothesis, we have created a recommender system for virtual community memberships in social networks.

LITERATURE REVIEW AND PREREQUISITES

Recommender systems: Personalized recommender systems have been in use since 1970 (Zhang and Koren, 2007); and they have been studied in two general

Corresponding Author: Hamid Alinejad-Rokny, Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

categories: content-based filtering systems and collaborative filtering systems (McCarthy *et al*., 2006; Chen *et al*., 2008; Garcia *et al*., 2009; Jameson, 2004).

Content-based filtering recommender systems, analyze item specifications to find the items that suit user preferences. In this method, the user profiles include features and items chosen by the user in the past; Infofinder, Newsweeder and News Dude are examples of such systems (Chen and Chen, 2005). Consequently, item descriptions and those of the user in his profile are the basis for offering recommendations in this system. The content-based filtering systems have two main weaknesses: they only recommend items similar to those chosen in the past; also, the system must extract meaningful and useful features from the content.

Collaborative filtering systems give recommendations based on the retrieved information from users who shared similar interests and preferences in the past (Zhang and Koren, 2007). Using this method, the system could use the feedbacks received from other similar users to give recommendations. The main advantage of this method over the previous one is that, in this method, the community of users could give their opinions about certain items, and also provide feedbacks and ratings; as a result, there is a chance for quite new items to be recommended to the user (Sobecki, 2006). The main purpose of this method is to offer recommendations that, considering profiles similarity, fall in the same category; examples include Ringo and Siteseer. Video Data and Personalized Television System are also the two systems that give new recommendations to users based on their previous preferences (actors, screenplay, movie and etc.) (Chen and Chen, 2005). considering the nature of this method, the system provides the system will provide weak and non-transparent predictions in case the number of users should be too few, (Cold Star) (Sobecki, 2006; Debnath, 2008). Other drawbacks of this method include First Rater, Sparsity and Popularity Bias (Debnath, 2008). Memory-based, Model-based and hybrid are three types of collaborative filtering. Neighborhood based CF and item based/user based top-N recommendations are two methods of memory-based CF; likewise, model-based CF include clustering and classifying based techniques. In each of these methods, it is tried to overcome some of the shortcomings and limitations of collaborative filtering. Su and Khoshgoftaar describe all different types of CF methods and compare them (Su and Khoshgoftar, 2009).

Hybrid System, to overcome the before-mentioned weaknesses as well as improve the accuracy and the effect of recommendations on users, researchers have tried to make hybrid systems by combining and mixing different methods; Tapestry and Grouplens are examples of such hybrid systems. In these systems, users are allowed to comment and rate on the basis in which they are then categorized. Also, to provide better and more effective recommendations, users must complete their profiles and include item features that are important to them (Chen and Chen, 2005).

Conditional entropy: In information theory, entropy is a measure of uncertainty associated with a random variable. In this context, the term usually refers to the Shannon entropy (Gray, 2009). In another words the random variable X's entropy which is also called X's source, indicates the average of produced information by the X's source. As the event's probability decreases, its uncertainty would increases:

$$H(x) = -\sum_{X} P(X)\log p(x), \quad H(x) \geq 0 \tag{1}$$

P(x) is the probability mass function of X and sigma includes all discrete values of X. Regarding to entropy definition, the conditional entropy quantifies the remaining entropy of a random variable Y, given that the value of another random variable X is also known. It is referred to as the entropy of Y conditional on X, and is shown as H(Y|X) (Gray, 2009).

$$H(Y|X) = \sum_{x \in X, y \in Y} P(X,Y) \log \frac{p(x,y)}{p(x)} \tag{2}$$

Decision tree classification: Problems like recommender systems are solved based on classification and prediction. There are many methods for classification; the nearest neighbor, neural networks, regression and D-tree are examples of such methods. One of the common methods in classification is D-tree; D-tree is a tree structure like flowchart. Therefore, leaves represent classes and branches stand for combinations and conjunctions of features which make a class. In a D-tree structure, the predictions are explained with "if-then" rules. D-tree is used with various numeric and ordinal data types. Furthermore, this shows which fields or variables have a crucial importance in the final prediction and classification. The importance of the variable grows as its distance to the root decreases. Classification is a 2 step process; the first step is making the model with the train data set and the second step is applying the model and classifying the test date. A model's accuracy depends on the number of its correct predictions. In fact it is the percentage of the number of times that test instances are classified successfully.

Various algorithms are applicable for D-tree; C5.0, C4.5, CART and etc. are examples of such algorithms. These algorithms could create binary or multi-branch trees (Han and Kamber, 2006; Tan *et al*., 2005).

Association mining: Association analysis is the discovery of association rules illustrating attribute-value

discovery of association rules illustrating attribute-value conditions that occur frequently in a given set of data (Han and Kamber, 2006). $I = \{i_1, i_2, \ldots, i_n\}$ is a set of n binary attribute called items. $D = \{t_1, t_2, \ldots, t_n\}$ is a set of transactions called database. Each transaction t has a unique transaction ID and contains a subset of items in I. each $X \rightarrow Y$ rule is an implication where $X, Y \subseteq I$ and $X \cap Y = 0$ are defined. The set of X and Y items are called antecedent or left-hand-side (LHS) and consequent or right-hand-side (RHS) (Han and Kamber, 2006; Yeh *et al*., 2009).

Association rules could be extracted based on different parameters. In addition, we may select interesting rules out of extracted rules based on different criteria. Some of these evaluation metrics are objective such as support, lift, mutual information, confidence (Tan *et al.*, 2004, 2006), Jaccard and etc.; some other evaluation metrics are subjective. Supporters of subjective measures believe that objective measures alone, are not capable of dealing with all the complexities of the knowledge discovery process. In fact a rule and pattern that is considered interesting based on the objective criteria, may not be considered so in an expert's opinion. Therefore, subjective criteria do not have a simple definition (Tan *et al*., 2006). Some objective criteria like lift and mutual information try to approach subjective criteria and select rules that are interesting to users.

We will explain the evaluation metrics used in this article. In order to have a high qualified association rule, we need a threshold for the least certain value. The rules that have a more value than threshold are called strong rules.

Support: The association rule's support is the number or percentage of transactions in a whole date set which includes both antecedent and consequent. For instance, we define the support for the rule $A \rightarrow B$ as Eq. (3). Where M is the number of transactions including A and B and N is number of all the transactions:

$$Support(A \rightarrow B) = \frac{m}{N} \tag{3}$$

Confidence: This parameter indicates the dependency of a special item on the other one and it is considered as an indicator for measuring a rule's power. The $A \rightarrow B$ rule's confidence is defined as below:

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \tag{4}$$

Lift: The lift measure is calculated based on Eq. (5) and it indicates A and B dependency:

$$Lift = \frac{Confidence(B|A)}{Support(B)} \tag{5}$$

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{6}$$

Mutual information: In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables. Formally, the mutual information of two discrete random variables X and Y is defined as (Tan *et al*., 2004).

where, P(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively. Therefore, when an association rule includes more mutual information comparing to the others, it is more interesting and more valuable.

RECOMMENDATION METHODOLOGY

This section explains research data set, extracting user profiles, identifying main users of groups and creating a personalized recommender system. Different steps required in a general framework for social network recommender systems are shown in Fig. 1. Here we explain five common group recommending techniques.

Main research data set: The data set used in this research is from Parsi-Yar's database, a Persian social network. This covers a 5-year and half activity duration. This data contains user information, groups' members and user interactions with other members and groups. Parsi-Yar has more than 3300 groups in 19 different categories; they include: youth, health, religion, politics, sports, entertainment and etc. Due to many changes applied to the site, the data in the database lack sufficient integrity and much of information, saved in text format, needed to be classified. Consequently, to study and analyze the information, under complicated and time-consuming processes, the un-structed data, were converted into structed and appropriate ones. In this paper the main data set was created manually.

The main two phases of data preparation in our research are data cleaning and data transformation (Han and Kamber, 2006). Generally dealing with null values, detection and emission of outlier data, integrated schema and dealing with data redundancies are some of the most important tasks in the data cleaning phase. Likewise, the important tasks in the data transformation phase are normalization, extraction and creation of new attributes and generalization of data values. Here we have tried to prepare efficient data by practicing all above-mentioned phases. It is not possible to mention all preprocessing steps in this study, but we will illustrate some points here.

Since our data are recorded in English, formal Persian, spoken Persian or a transliteration of one of them

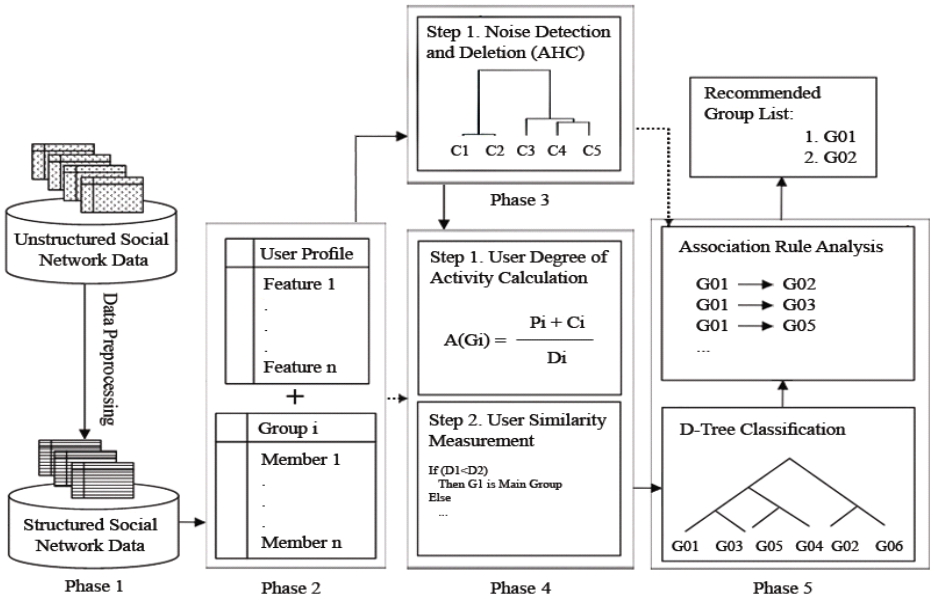

Fig. 1: The main core of the group recommender system framework consists of five phases: 1.Creation of the main data set and pre-processing, (2). Identification of user profiles and groups' members, (3). Identification of the main members of groups based on users features,(4). Identification of the main group for each user and (5). Creation of the recommender system and providing the list of recommendations. In addition, the output data of each phase makes the input for the next one; the dotted arrows represent the lateral route of data

in the other language, through a creative method based on repeated and self-learning data mining techniques, we detected all non-Persian data and converted them into their formal Persian equivalences.

We used two different methods to classify some features of user information: three-level tree structure for the job and expertise attributes; and one-level classification for all other attributes. To classify some of the features, experts and reference websites were consulted; in some cases, we managed to convert un-structed data to structed ones by interviewing people and doing surveys. For instance, in some cases, users had entered a specific singer for "your favorite music genre" field; in such cases we identified different music genres and figured the special music genres of different singers and finally extracted the favorite music genre of each user.

We designated each feature a new "undefined" category. Furthermore, since assigning any values to features with missing values would affect the reliability of the research and consequently the final decision making, to maintain accuracy and reliability of data, users with missing values were tagged undefined.

In some rare cases, we managed to assign correct values these missing value features by considering those of other features. For instance, a user had *basketball* as "hobbies" in his profile but did not record any values for the "favorite sports" feature; in this case we recorded basketball as his favorite sport. Moreover, to maintain the original data and reliability as much as possible, we created an "others" category to which all value classes with too few members were assigned.

In some cases, features such as age, date of birth, height and weight, *outlier values* were identified and were assigned values based on the values of other features of that user or those of other members. In the end, 54.49% of each user profile was filled. We standardized all features values between 0 and 1 by Min-Max normalization method.

Identifying user profiles and groups: Data reduction is another important process in data preprocessing. After creating an appropriate primary data set, to make user profiles the basis of our analysis, we managed to identify 49 features for each user based on their profiles as well as their links and interactions with other users and groups in the social network. Some of these features include: sexuality, language, level of education, height, weight, age, date of birth, city, marital status, job, expertise, favorite books, music, movies and sports, number of friends, date of membership, number of joined groups, number of sent posts in the group, number of replies to the posts in the group and etc.

Out of more than 3300 groups in our data set, only 27 groups had more than 100 users; out of which we selected 15 most active and popular groups as the target (Table 1);

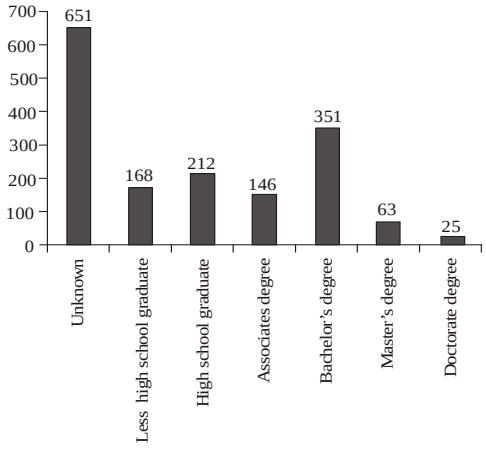

Fig. 2: Diagram of sexuality of users in target groups: Male sexuality make the majority of groups' members

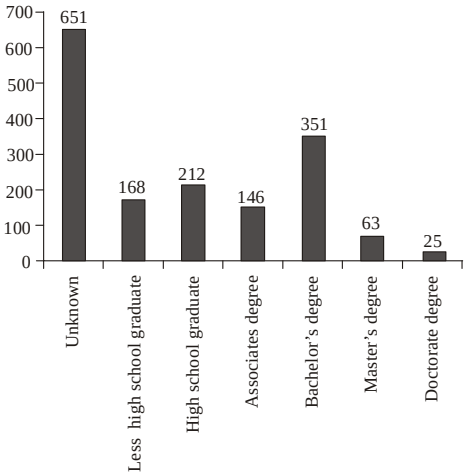

Fig. 3: Diagram of educational level of members of target groups: The educational level of the majority of users is undergraduate or below

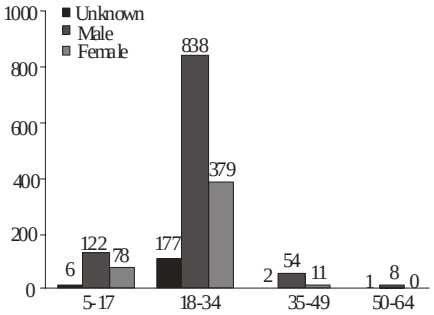

Fig. 4: Diagram of age and sexuality of target groups members

users had multi-memberships; for example, a user was member of G04, G03, G01 and G07. On average, users have memberships in two groups.

Figure 2 and 3 illustrate the educational status as well as the sexuality of members of target groups accordingly. The sexuality and age of members of the target groups are shown in Fig. 4. As illustrated, in our social network the majority of groups' members are male.

Identifying main users of groups: We believe that, despite differences, users in a certain group share similar characteristics and personalities. Nonetheless, there are members in every group that differ from the rest of the group; they are called noises. Experts believe that they, the noises, are users who have joined a group unknowingly or out of curiosity. To identify the main users, we use hierarchical clustering; and to determine the similarities between users we apply Euclidean Distance Eq. (7). In which d stands for similarity measure (distance) between nodes s and r; N is the number of dimensions or features of user profile and x is value of a feature in r and s. To use Euclidean Distance, the values should be in a standard limit (Han and Kamber, 2006).

$$d_{rs} = \sqrt{\sum_{i=1}^{N} (x^r - x_s)^2} \quad (7)$$

Hierarchical Cluster Analysis (HCA): Primary goal of cluster analysis is to classify objects; therefore, all similar objects must fall in the same cluster (Hair *et al*., 1998). The resulting clusters of objects should show high internal homogeneity and high external heterogeneity (Hair *et al*., 1998). Clustering analysis may reveal associations and structures in data that were not evident in the first place; nevertheless, once found, they are predictable and useful (Yeh *et al*., 2009). HCA is a general view for cluster analysis. An important step in this method is repetition of calculations of the distances between objects and between clusters to which those objects belong. The output is presented as a dendrogram. The frequently used algorithms for hierarchical clustering are: centroid method, average linkage, complete linkage, single linkage and Ward method. The main difference of these methods lies in the way the distance between clusters is calculated (Hair *et al*., 1998; Yeh *et al*., 2009).

We used Ward algorithm in this research which is available in a Microsoft Excel add-Ins named XLSTAT; Sum of the Squared Error (SSE) internal measurement was used to evaluate the effectiveness of clustering Eq. (8). The SSE for a cluster is the sum of distances between its internal members and its centroid (Tan *et al*., 2005).

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \sum_{j=1}^{n} \left(m_j i - x_j \right)^2 \quad (8)$$

K is the number of clusters, C_i stands for cluster members, x is one of the members, n is the number of attributes for

Table 1: 15 most popular groups' information

Group ID	Category	Initial group size	No. of single-memberships	Group ID	Category	Initial group size	No. of single-memberships
G01	Social	373	225	G02	Social	158	87
G02	Social	219	98	G04	Social	133	54
G05	Sport	140	88	G06	Morality and spirituality	212	120
G07	Youth	207	123	G08	Morality and spirituality	151	87
G09	Revolution	137	85	G10	Religion	154	90
G11	Literature	276	188	G12	Entertainment	134	80
G13	Social	131	70	G14	Social	211	112
G15	Entertainment	151	109				

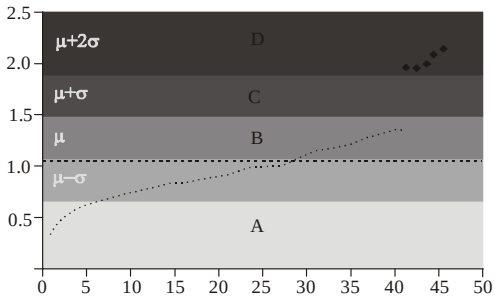

Fig. 5: Normal distribution of distances of users in a certain cluster from its centroid: The vertical axis shows the distance from the cluster centroid. In addition, the existing members in a cluster are drawn on the horizontal axis

a member in the cluster, m_j^i is the jth attribute of centroid member of cluster C_i , and x_j is the jth attribute of the member in the cluster (Tan *et al*., 2005).

Detection and deletion of noises: We used Normal (Gaussian) distribution in each group and each cluster to identify the noises. Members of one of the G03 clusters are shown in Fig. 5. In each cluster, users are illustrated based on their distance from the cluster centroid. Users who are placed in area A have the greatest similarities with the cluster. Users who are placed in area B are those who have an acceptable similarity with the cluster centroid and which cannot be marked as noise. Nonetheless, users in area C might be considered as noise or a main user, depending on their values. Finally, users in area D are known as noise and deleted. As a result, the remaining members of the group will be more homogeneous; and therefore, decisions will not be influenced by noise spots.

Identifying the main group of each user: Users in a social network could join and be active in more than one group and community. To provide a unified model in our method, we need to identify the main group for each user. lti-memberships lead to decisions that will ultimately be more based on probabilities than facts. Therefore, by identifying the main group of each user and ignoring multi-memberships, the main users of each group will better be identified in this step of our modeling. After the first step (identifying main users of groups), multi-memberships were reduced to 31.63%.

We define user activity degree to figure out the level of activity for each user Eq. (9); and in case a user is more active in a certain group, he will be identified as a member of that group only and his other memberships in other groups will be ignored by the system in this phase.

$$A(G_i) = \frac{P_i + C_i}{D_i} \quad (9)$$

$A(G_i)$ is the activity degree of user in *i*th group, P_i is the number of posts sent by the user, C_i is the number of replies and comments given by the user to the posts in the group and D_i is the duration of membership of the user. Therefore, any group with the highest degree of activity is considered as the user's main group. After this phase, multi-memberships were reduced to 22.6%.

Finally, since not all users are active in group and we cannot use activity degree to determine single-memberships, we took into account the similarity of user to other members of a certain group. We considered the distance between each user and the cluster centroid; and the group whose user had the least distance, was regarded as the main group for that user. In other words, each user was assigned a group with which it had the most possible similarity. After all, 1616 users belonged to one single group. Table 1 shows the number of single-memberships of 15 groups after assigning a main group to each.

Creating recommender system: We used D-tree to predict user main membership in a group. Finally, regarding association rules and evaluation metrics, we recommended group lists of two that were related to user features. Since 31.63% of users had multi-memberships and others were members of only one group and users in our social network were members of two groups on average, we recommended a list of two groups for each user; Of course more groups could also be recommended. As a result, user multi-memberships in groups, which were ignored in identification of a main group for each user, will again be influential. In this step we consider multi-memberships to offer recommendations.

When for instance, D-tree predicts G01 for one user, we consider the rules whose first side (antecedent) is G01;

for example: four consequents: G05, G03, G02 and G10. In addition, the group which is more suitable based on metrics of association rules evaluation will be recommended to the user along with G01; for example G02. As a result, we recommend G01 and G02 to the user.

D-tree to predict main group: D-tree is used to recommend a main group to the user. Furthermore, for our recommendations to be related to user's main features, we used supervised entropy (conditional entropy) (Han and Kamber, 2006; Gray, 2009). Considering Eq. (2), Y represents our target groups which includes 15 distinct values; and X stands for 49 user features, each of which includes different values. Out of 49 primary features, 20 were identified as less related or unrelated; and only the 29 remaining features of user profiles were used in D-Tree classification. Features like favorite season, eye color, skin color, hair color, level of education and etc. are example of the 20 unrelated or less related features.

Our model created by a binary D-tree and we used C5.0 algorithm and boosting (Freund and Schapire, 1999). *Boosting* is a general technique for improving classifier accuracy (Han and Kamber, 2006) and in fact, it is a machine learning meta-algorithm for performing supervised learning.

The data set applied to this step includes user profiles and requires membership in one group for each user.

Association rules for determining groups related to the main group: This step is top-N recommendations. Association rules were used to provide a maximum number of two group recommendations to each user. Here, groups are considered as data items and each user's multi-memberships are considered as a transaction.

In this research, we select interesting rules based on lift, confidence and mutual information metrics. The data set applied to this step includes user multi-memberships in groups, without considering noise users.

Other methods for group recommendations: Here, we will introduce five methods based n traditional and common recommender systems techniques such as CB and CF methods in two categories.

Non-personalized recommendations:

Popular groups with more users: in this technique, the two popular groups which have the most number of users are recommended. This method is non-personalized and recommends the same groups to all users.

Personalized recommendations:

Collaborative filtering: In this method, users with similar profiles are put into similar cluster; in each cluster, the two groups that have the most number of users will be recommended to the users in that category. In our social network, there is no option for users to give feedbacks to groups or rate them. Therefore, we consider the membership of users as their rating; One's membership means rating 1 and otherwise is regarded as rating 0. HAC is used to classify users with similar profiles. The limitation of this method for users who have just joined the social network or users who do not have any group membership is that either there is no recommendation or the offered recommendation is not a valid one.

- **First CF method:** in this technique users are classified based on their similarity in groups' memberships.
- **Second CF method:** in this hybrid technique users are classified based on their similarity in group memberships as well as personal features. In this technique, unlike first CF method, personal features of users as well as their history of group membership consist user profiles.

Recommendation based on users' friends: This method is based on the concept of graphs and neighborhood based CF. In this technique, the two groups which have the most popularity between users' friend are recommended to that user. The limitation for this method is that no recommendation will be provided for users who have just joined the social network or those who do not have any connections with other members.

Recommendation based on association rules: In this top-N method, the relationship between groups and user membership in groups are focused. For each group, by using association rules and considering user memberships, the two much closely related groups are found. By the relationship between A and B we mean that, for example, most users who are members of group B are also members of group A; therefore, group B and group A are related. As a result, when a user is a member of group B, we recommend group A. The limitation for this method is that users who are not members of any groups or who have recently joined the social network will not receive any recommendations.

Recommendation evaluation metrics: Two principal metrics applied in evaluation of recommender systems are coverage Eq. (10) and precision Eq. (11). Coverage is about how well the system covers user demands and precision is the ratio of useful recommendations to all recommended items. Therefore, when there are a scarce number of wrong recommendations, precision will improve (Badrul *et al*., 2001):

$$\text{Coverag} = (|\text{Recommended Items}| \cap |\text{Favorite It} / |\text{Favorite Items}|) \quad (10)$$

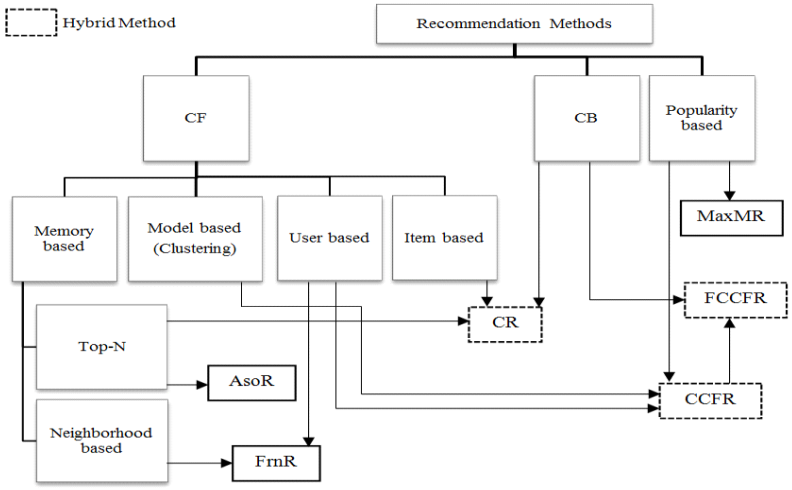

Fig. 6: Illustration of the techniques presented in this study

$$\text{.Precision} = \left(\left|\text{Recommended Items} \cap \text{Favorite It}\right| / \left|\text{Recommended Items}\right|\right) \quad (11)$$

There are two general approaches for evaluation. The first one is more of a qualitative approach in which results are shown to some sample users; the measure of coverage and precision criteria are acquired from user comments to be used in comparing this system with other recommender systems. The second approach is a more strictly one. First, a set of favorite items are collected; a part of this collection is given to the recommender system as the training data and the remaining items are used for system evaluation. The second approach, compared with the first one, is more precise and strict.

EXPERIMENTAL RESULTS

The methods introduced in this paper are shown in . Fig. 6. We considered different recommendation techniques:

- The technique presented by this paper (CR)
- Popular groups with more users (MaxMR)
- First CF method (CCFR)
- Second CF method (HCCFR)
- Recommendation based on users' friends (FrnR)
- Recommendation based on association rules (AsoR)

In this study we used the second approach to evaluate the recommender systems. To do so, 75% of data were used in the training phase and the remaining 25% were used in the test phase and sampling was randomly done.

In CR, 29 user features were selected using supervised entropy to be used in D-tree. In addition, MaxMR, CCFR, HCCFR, FrnR and AsoR are based on data from which noise members are detected and deleted. The resulting D-tree precision was 75.57% for CR method. The accuracy of D-tree prediction model that is presented in Table 2 is based on the results of repetitive test. As shown in Table 2, the CCFR model accuracy that has binary class value of inputs is better than ones have multiple class values.

Table 2: Accuracy of D-tree classification

Method	Number of input	Boosting usage	Accuracy (%)
CR	29	Yes	75.57
CCFR	15	No	96.83
FCCFR	64	Yes	83.09

Table 3: Association mining details

Rule extracting measure	Minimum rule confidence (%)	Minimum antecedent support (%)	No. of rules
Rule confidence	15	5	66
Information difference	15	5	64

By setting support threshold to 5% and rule confidence threshold to 15%, all possible association rules were extracted for CR and AsoR methods (Table 3).

The average of repetitive tests (20 tries) is illustrated in Table 4. As shown in Table 4, results have been presented based on different techniques. In addition, for CR association rules are evaluated by different metrics of extracting and selecting interesting rules.

In CR method, rules extracted by confidence rule and information difference, have brought about recommendations with closely identical precision and coverage. Also, the *confidence* subjective criteria have had better results in selecting rules. For its minor differences with the results of *mutual information*, both metrics could be considered as acceptable.

Table 4: Average of coverage and precision ($1 \leq N \leq 15$)

Method	Metrics for evaluating rule extraction	Metrics for rule selection	Number of user's favorite group	Precision	Coverage
CR	Confidence rule	Confidence	N	45.18%	66.27%
	Information difference	Confidence	N	45.18%	66.27%
	Confidence rule	Lift	N	43.84%	64.97%
	Information difference	Lift	N	43.84%	64.97%
	Confidence rule	Mutual information	N	44.29%	65.42%
	Information difference	Mutual information	N	44.29%	65.42%
MaxMR	-	-	N	15.02%	20.30%
CCFR	-	-	N	20.41%	40.21%
FCCFR	-	-	N	24.20%	44.29%
FrnR	-	-	N	25.64%	31.20%
AsoR	Confidence rule	Confidence	N	11.60%	7.31%

In the extracted association rules, users whose main group is about social subjects, receive a group recommendation around the same topic. In addition, users who their main group is morality and spirituality, their second group is also morality and spirituality. Furthermore, users interested in groups around the topics of revolution, are also interested in religion and their second group is of that category. Accordingly, other users who are interested in groups with topics around sports, youth, literature and entertainment are offered a group around social topics.

Features closer to the tree root have more importance in the decision making and classification; therefore, year of membership, favorite movie, weight, date of birth and favorite book are more effective and more important features in predicting the main group for users.

DISCUSSION AND CONCLUSION

Social networks have recently gained a lot of importance and attracted many users. The multiplicity of virtual groups and communities confuses users and makes it very difficult for them to join groups that perfectly suit their character and satisfy their needs; this eventually may lead to user boredom. The hybrid group recommender system explained in this study was a solution for such problems by offering customized recommendations based on user features. Regarding the similarities between members of a group, this system identifies each user's main group and applies D-tree classification method to recommend one as the main group with the highest priority. In addition, based on association rules, the system identifies one other group as related to user's main group and recommends it to him.

Compared with the common methods and unlike CF and CB recommender systems, the techniques introduced in this study provide better results and lack their limitations. This technique, regardless of users' activity in the network, could offer recommendations even to users who do not have any connections or memberships in groups in the past; and this could be a great advantage over the previous techniques; moreover, this system could also recommend more than two groups. As a result, by this hybrid method, we have managed to solve some limitations of other recommender systems.

In the others method also, the FrnR method had a better precision which is a natural phenomenon due to the fact that friends influence their friends (Sinha and Swearingen, 2001). Regards to our evaluation approach, in AsoR method, users who are member of just one group do not receive any helpful suggestions. Therefore its precision and coverage is less than other methods. Although the accuracy of D-tree prediction in CCFR and FCCFR models are more than the hybrid CR method, the precision and coverage of CR method is more than those. In addition, there is no need to all user information would be completed, although the hybrid CR method is based on content; in this research, CR model is constructed just by knowing 54.49% of user information.

The framework in this study could also have other applications; for example where recommended items do not have any special features or attributes, or where it is needed to apply text mining techniques for identification and extraction features. Moreover, in some cases which it needs to offer recommendations to a group of people, step 3 of our system is useful for identifying and specifying group features. Using this method is appropriate when user-item matrix is sparse and CF methods do not deliver good results.

In the future studies, we mean to give recommendations by considering user relations based on the graph theory and combining other CF methods with CB.

REFERENCES

Adamic, L.A., O. Buyukkokten and E. Adar, 2003. A social network caught in the web. First Monday, 8(6).

Badrul, M., G. Sarwar, J. Karypis, A. Konstan and J. Riedl, 2001. Item-based collaborative filtering recommendation algorithms. Proc. WWW' 01, The 10th International conference on World Wide Web, pp: 285-295.

Chen, H.C. and A.L.P. Chen, 2005. A music recommendation system based on music and user grouping. J. Intelligent Inf. Syst., 24(2): 113-132.

Chen, Y. L., L.C. Cheng and C.N. Chuang, 2008. A group recommendation system with consideration of interactions among group members. Expert systems with applications. Int. J., 34(3).

Debnath, S., 2008. Machine Learning Based Recommendation System. Thesis submitted in partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering, Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur.

Freund, Y. and R.E. Schapire, 1999. A Short Introduction to Boosting. J. Japanse Soc. Artificial Intelligence, 14(5): 771-780.

Garcia, L., E. Sebastia, C. Onaindia and C. Guzman, 2009. A group recommender system for tourist activities: Proc. EC-Web. Lecture Notes Computer Science E-Commerce Web Technologies, 5692: 26-37.

Gray, R.M., 2009. Entropy and Information Theory. Springer-Verlag, New York.

Hair, J., R. Anderson, R. Tatham and W. Black, 1998. Multivariate Data Analysis. 5th Edn., Prentice Hall, NJ.

Han, J. and M. Kamber, 2006. Data Mining. Concepts and Techniques, Elsevier.

Jameson, A., 2004. More Than the Sum of Its Members: Challenges for Group Recommender Systems. Proceeding of International Working Conference on Advanced Visual Interfaces, pp: 48-54.

McCarthy, K., M. Salamó, L. Coyle, L. McGinty, B. Smyth and P. Nixon, 2006. Group Recommender Systems: A Critiquing Based Approach. Proc. IUI '06, The 11th International Conference on Intelligent User Interfaces, pp: 267-269.

Sinha, R. and K. Swearingen, 2001. Comparing recommendations made by online systems and friends. In Paper presented at the second DELOS network of excellence workshop on personalization and recommender systems in digital libraries.

Sobecki, J., 2006. Implementation of Web-based Recommender Systems Using Hybrid Methods. Int. J. Comp. Sci. Appl. (IJCSA), 3(3): 52-64.

Su, X. and T.M. Khoshgoftar, 2009. A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence, pp: 1-19.

Tan, P.N., M. Steinbach and V. Kumar, 2005. Introduction to Data Mining Addison. Wesley.

Tan, P.N., M. Steinback and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley.

Tan, P.N., V. Kumar and J. Srivastava, 2004. Selecting the right objective measure for association analysis. Inf. Syst., 29(4): 293-313.

Yeh, C., C.H. Lien, T.M. Ting and C.H. Liu, 2009. Application of web mining for marketing of online bookstores. Expert Syst. Appl., 36(8): 11249-11256.

Zhang, Y. and J. Koren, 2007. Efficient Bayesian Hierarchical User Modeling for Recommendation Systems. Proceeding of SIGIR '07. The 30th Annual International Acm Sigir Conference on Research and Development in Information Retrieval.