Group Recommendation System for Facebook

Enkh-Amgalan Baatarjav, Santi Phithakkitnukoon, and Ram Dantu

Department of Computer Science and Engineering
University of North Texas, Denton, Texas, 76203, USA
{eb0050,santi,rdantu}@unt.edu

Abstract. Online social networking has become a part of our everyday lives, and one of the popular online social network (SN) sites on the Internet is Facebook, where users communicate with their friends, join to groups, create groups, play games, and make friends around the world. Also, the vast number of groups are created for different causes and beliefs. However, overwhelming number of groups in one category causes difficulties for users to select a right group to join. To solve this problem, we introduce group recommendation system (GRS) using combination of hierarchical clustering technique and decision tree. We believe that Facebook SN groups can be identified based on their members' profiles. Number of experiment results showed that GRS can make 73% accurate recommendation.

Keywords: Social network, recommendation system, decision tree.

1 Introduction

Face-to-face, voice, email, and video communications are traditional medium of interaction between friends, family, and relatives. The traditional medium takes place when two parties had already shared some form of common value: interest, region, family bond, trust, or knowledge of each other. Although, on online social network (SN) two parties initiate communication without the common values between them, they still can freely share their personal information with each other [1]. In the virtual world, joining or creating groups and making friends are a click of a button, which makes online social networking sites, such as Friendster, MySpace, Hi5, and Facebook more and more popular and diverse each day [14]. Therefore, online SN's advantages are user friendliness and flexible in cyberspace where users can communicate with others and create and join groups as their wishes.

Even though flexibility of online SN brings diversity in cyberspace, it can also lead to uncertainty. We took University of North Texas (UNT) SN as a sample for our research. There are 10 main group types, such as business, common interest, entertainment & arts, geography, music, etc. Six of them have over 500 groups, and four of them have range between 61 and 354 groups in each. It is overwhelming to find a group that fits a user's personality. Our study concentrates on identifying inherent groups' characteristics on SN, so that we develop group recommendation system (GRS) to help the user to select the most suitable group to join.

R. Meersman, Z. Tari, and P. Herrero (Eds.): OTM 2008 Workshops, LNCS 5333, pp. 211–219, 2008.

Groups were created to support and discuss causes, beliefs, fun activities, sports, science, and technology. In addition, some of the groups have absolutely no meaningful purpose, but just for fun. Our research shows that the groups are self-organized, such that users with similar characteristics, which distinguishes one group from others. The members' characteristics are their profile features, such as time zone, age, gender, religion, political view, etc, so members of the group have some contributions to their group identity. The group members' characteristics shape characteristic of the group.

Main Contribution: In this paper, we present Group Recommendation System (GRS) to classify social network groups (SNGs). Even though groups consist of members with different characteristics and behaviors, which can be defined by their profile features, as their group size grow, they tend to attract people with similar characteristics [13]. To make accurate group recommendation, we used hierarchical clustering to remove members whose characteristics are not quite relevant with majority in the group. After removing noise in each group, decision tree is built as the engine of our GRS. In this paper, we show how decision tree can be applied not only to classifying SNGs, but also used to find value of features that distinguish one group from another. GRS can be a solution to online SN problem with the overwhelming number of groups are created on SN sites because anyone can create groups. Having too many groups in one particular type can bring concern on how to find a group that has members who share common values with you. We believe if more and more members share common values, the group will grow in size and have better relationship. Thus, GRS can be a solution to many SNG issues.

The rest of the paper is organized as follows. In Section 2, we discuss related work done on social network. In Section 3, we describe the architecture and framework of GRS. Section 4 presents the performance of GRS. The paper is concluded with summary and an outlook on future work.

2 Related Work

There has been an extensive number of research efforts focused around modeling individual and group behaviors and structure, but due to its vastness we restrict here to providing only a sample of related research projects. Many researches on social networking have been done in mathematics, physics, information science, and computer science based on properties, such as small-world, network transitivity or clustering, degree distributions, and density ([6],[7],[8],[10], and[11]).

From research in statistics, Hoff et al. [9] developed class models to find probability of a relationship of between parties, if positions of the parties are known on a network. Backstrom et al. [2] has done very interesting research on finding growth of network and tendency of an individual joining a group depends on structure of a group.

3 Methodology

In this section, we cover data collection process, noise removal using hierarchical clustering, and data analysis to construct decision tree. Figure 1 shows basic architecture of the group recommendation system (GRS). It consists of three components: i) profile feature extraction, ii) classification engine, and iii) final recommendation.

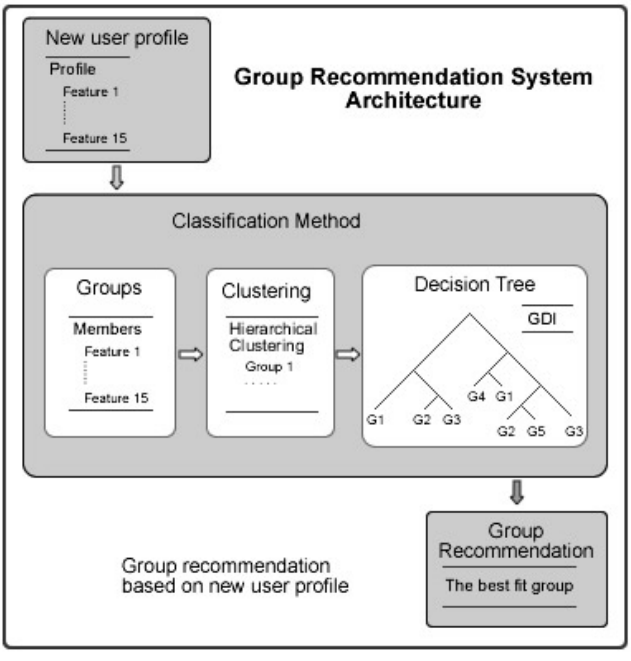

Fig. 1. Basic architecture of GRS, which consists of three major components including profile feature extraction, classification engine, and final recommendation

3.1 Facebook API

The dataset we used in this research was collected using Facebook Platform. Facebook launched its API to public in May 2007 to attract web application developers. The API is available in multiple programming languages: PHP, Java, Perl, Python, Ruby on, C, C++, etc. Since Facebook and Microsoft became partners, Microsoft has launched developer tools in its Visual Studio Express and Propfly. The Facebook Platform is REST-based interface that gives developers access to vast amount of users' profile information.

Using this interface, we had access to student accounts in which privacy setting was configured to allow access to its network (default setting). In our research we used University of North Texas (UNT) social network on Facebook. During this research we were able to access 1580 users' accounts. From the accounts, we collected users' profile information, friend connections, and groups where they belong to. For our analysis, we selected 17 groups from common interest groups on UNT SN. Table 1 shows detailed information of the groups.

3.2 Profile Features

The first step of group recommendation system is to analyze and to identify the features which capture the trend of a user in terms of its interest, social connection, basic information such as age, sex, wall count, notes count and many such features.

Table 1. Information of 17 common interest groups on UNT social network including their subtype categories, number of members, and description

Group	Subtype	Group Size	Description
G1	Friends	12	Friends group for one is going abroad
G2	Politic	169	Campaign for running student body
G3	Languages	10	Spanish learners
G4	Beliefs & causes	46	Campaign for homecoming king and queen
G5	Beauty	12	Wearing same pants everyday
G6	Beliefs & causes	41	Friends group
G7	Food & Drink	57	Lovers of Asian food restaurant
G8	Religion & Spirituality	42	Learning about God
G9	Age	22	Friends group
G10	Activities	40	People who play clarinets
G11	Sexuality	319	Against gay marriage
G12	Beliefs & causes	86	Friends group
G13	Sexuality	36	People who thinks fishnet is fetish
G14	Activities	179	People who dislike early morning classes
G15	Politics	195	Group for democrats
G16	Hobbies & Crafts	33	People who enjoys Half-Life (PC game)
G17	Politics	281	Not a Bush fan

We extracted 15 features to characterize a group member on Facebook: Time Zone - location of the member, Age, Gender, Relationship Status, Political View, Activities, Interest, Music, TV shows, Movies, Books, Affiliations - number of networks a member belongs to, Note counts - number of member's note for any visitors, Wall counts - visitor's note for member's page, Number of Fiends - number of friends in the group.

Based on analysis of 17 groups, we found some interesting results of differences between groups. Figure 2 illustrates gender ratio, age distribution, and political view in 17 groups. It is also useful to draw parallel attention between Table 1 and Fig. 2. G1 is a friend group, and majority of the members are Female, age between 20 and 24, and 33% don't share their political preference. Same 33% are moderate. These properties identify G1. Same way we can interpret all 17 groups. Female members are majority in G1 (friends group), G4 (campaign for homecoming king and queen), G7 (Asian food lovers), G10 (clarinet players), G13 (people who likes fishnet), and G17 (Not Bush fan). At same time, majority of G17 consider themselves as liberal. Fig. 2(b) shows that majority of all groups are members between age 20 and 24. Fig. 2(c) illustrates that majority of G3 (spanish learners), G5 (wearing same pants everyday), G7 (Asian food lovers), G8 (religions group), G10 (clarinet players), G12 (friends group), G16 (PC gamers) did reveal their political preference.

As we can see that using this property, we can construct a decision tree to make better group selection for Facebook users.

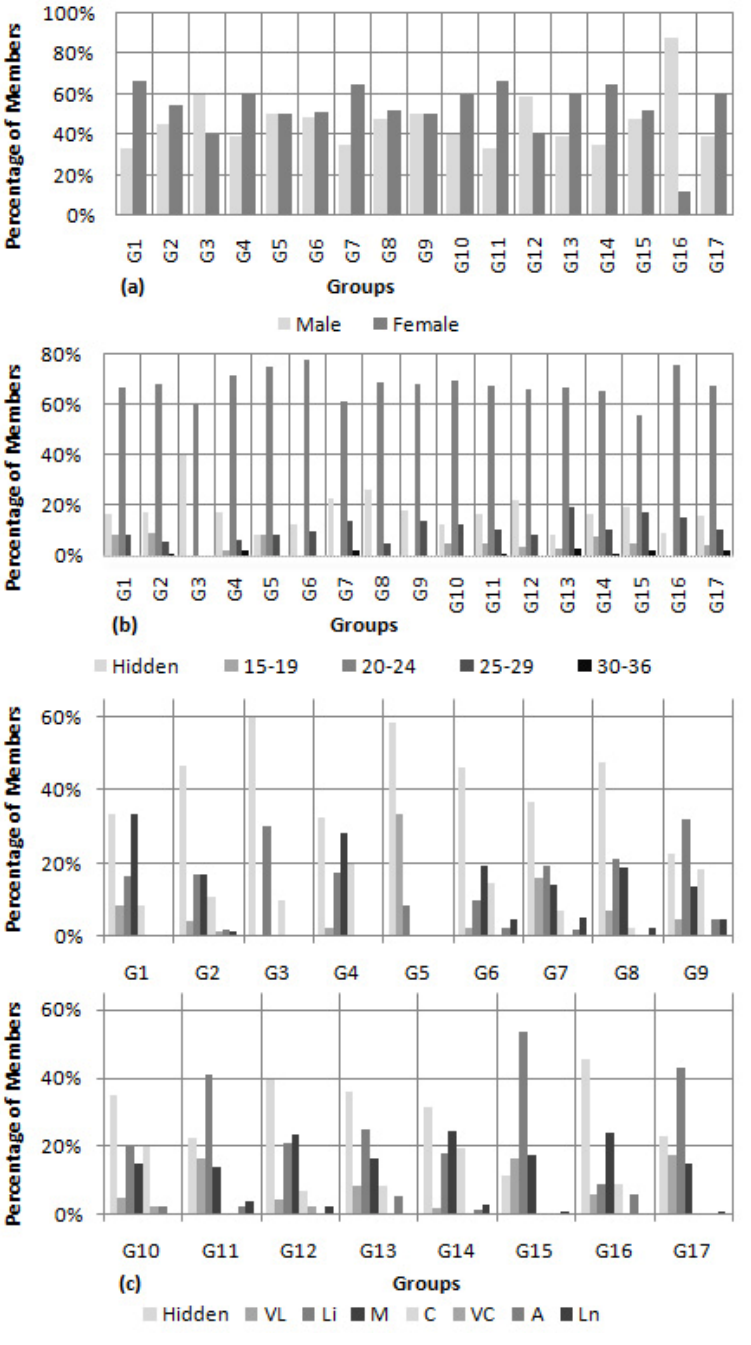

Fig. 2. (a) Gender ratio of each group. (b) Age distribution ranges of 15 to 19, 20 to 24, 25 to 29, and 30 to 36. (c) political preference distribution of the following very liberal (VL), liberal (Li), moderate (M), conservative (C), very conservative (VC), apathetic (A), and libertarian (Ln).

3.3 Similarity Inference

One of the frequently used techniques to find similarity between nodes in multidimensional space is hierarchical clustering analysis. To infer similarity between members, we use Euclidian distance [12].

Clustering takes place in the following steps for each group: i) normalizing data (each feature value = [0, 1]), ii) computing a distance matrix to calculate similarities among all pairs of members based on Eq. (1), iii) using unweighted pair-group method using arithmetic averages (UPGMA) on distance matrix to generate hierarchical cluster tree as given by Eq. (2).

$$d_{rs} = \sqrt{\sum_{i=1}^{N} (x_r - x_s)^2}\,, \tag{1}$$

where d is the similarity between nodes r and s, N is number of dimensions or number of profile–features, and x is value at a given dimension.

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj})\,, \tag{2}$$

where n_r is number of cluster in r, n_s is number of cluster in s, x_{ri} is the ith object in cluster r, and x_{si} is the ith object in cluster s. The Eq. (2) finds average distance between all pairs in given two clusters s and r.

Next step is to calculate clustering coefficient to find the cutoff point such that noise can be reduced. In Section 3.4 shows finding clustering coefficient.

3.4 Clustering Coefficient

Each group has a unique characteristic, which differentiates it from others, yet some members within the same group may have different profiles. As these differences grow to some extent, these members emerge as an inevitable "noise" for clustering.

To detect and mitigate this noise thus the group is strongly characterized by core members who establish innermost part of the group, we introduce the *clustering coefficient (C)*, which is given by Eq. (3).

$$C = \frac{N_{R_i}}{R_i}\,, \tag{3}$$

where R_i is the normalized Euclidean distance from the center of member i, given by Eq. (4) hence $R_i = [0, 1]$, and N_k is the normalized number of members within distance k from the center, given by Eq. (5) and hence $N_k = [0, 1]$.

$$R_i = \frac{r_i}{\max_j (r_j)}\,, \tag{4}$$

where r_i is the distance from the center of member i and $i = \{1, 2, 3, \ldots, M\}$.

$$N_k = \frac{n_k}{M}\,, \tag{5}$$

where n_k is the number of members within distance k from the center, and M is the total number of members in the group.

To reduce the noise in the group, we retain only members whose distances from the center are less than and equal to R^x as shown in Fig. 3, where R^x is the distance at which clustering coefficient reaches the maximum.

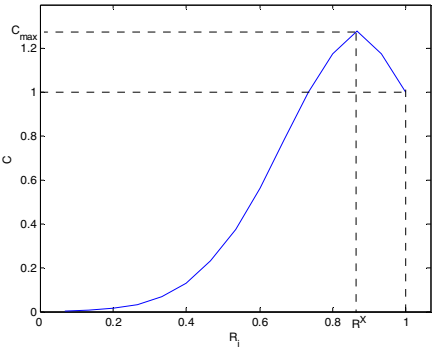

Fig. 3. An example of C-vs-R_i plot for finding R^x. This plot illustrates the cutoff distance R^x which is the corresponding distance from the center at which the clustering coefficient is maximum and gradually decreasing to 1 at $R_i = 1$. As the clustering coefficient starts to decrease, the sparseness of the outer circular members increases more rapidly since the denominator starts to dominate (greater than numerator) according to Eq. (3). Hence the outer circular members or members who have distance from the center greater than R^x to be considered as noise and removed.

3.5 Decision Tree

The nature of group recommendation system (GRS) is classification type problem. Based on a user's profile features, GRS finds the most suitable groups for a user. One solution to classification type problem is decision tree algorithm, based on binary recursive partitioning. There are number of splitting rules: Gini, Twoing, and Deviance [3]. To find better result we integrated each of splitting rule to GRS. However, test showed no significant improvement in accuracy, which means that final tree does not depend on what splitting rule is used to construct the tree [3]. The main goal of these splitting algorithms is to find the best split of data with maximum homogeneity on each side. Each recursive iteration purifies data until the algorithm reaches to terminal nodes (classes).

Binary tree consists of parent node t_p and child nodes of t_l and t_r. To define maximum homogeneity of child node, we introduce impurity function $i(t)$, so maximum homogeneity of t_l and t_r nodes is equal to the maximum change in impurity function $\Delta i(t)$ (given by Eq. (6)), which shows that splitting rule go through all variable values to find the best split question $x_i \leq x_j^R$, so that maximum $\Delta i(t)$ is found.

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r), \tag{6}$$

where P_l and P_r are probabilities of left and right nodes, respectively. Thus, maximum impurity is solved on each recursion step and given by Eq. (7).

$$\max_{x_j \leq x_j^R, j=1\ldots M} [i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r)], \tag{7}$$

where x_j is variable j, x_j^R is the best possible variable x_j to split, M is number of variables.

4 Result

In this research we developed group recommendation system (GRS) using hierarchical construct and decision trees. To evaluate the performance of GRS, we used 50% of data for training and other 50% for testing. For testing, we selected labeled members and clustered those using GRS. Accuracy rate is measured by the ratio of correct clustered members to total testing members. Figure 4 compares accuracy of GRS with clustering and without clustering for noise removal. Average accuracy without clustering was 64%. Meanwhile, after removing noise from each group using clustering coefficient method, average accuracy improved to 73%. In other words, average accuracy improved by 9%. In addition, 32% of 1580 members or 343 members were found to be noise and eliminated.

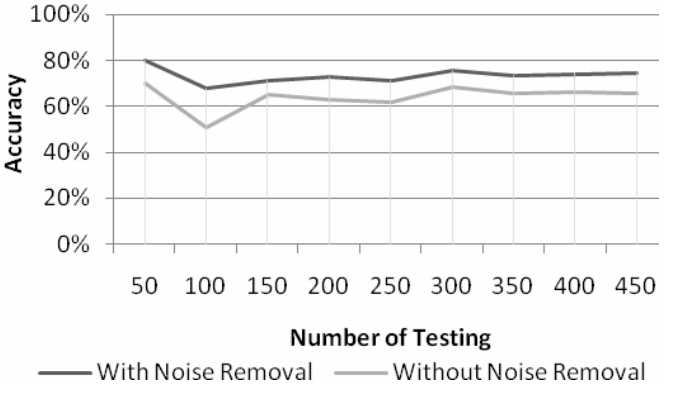

Fig. 4. Accuracy comparison of GRS with and without clustering where the accuracy is improved by 9% with clustering

5 Conclusion and Future Work

It is challenging to find a suitable group to join on SN, especially networks as big as MySpace and Facebook. Until now, online social networking has no sign of slowing down. While Facebook has 42 million users as of October 2007, there are 67 million active users as of February 2008. It has been doubling its size in every six months.

To improve quality of service for Facebook users, we developed GRS to find the most suitable group to join by matching users' profiles with groups' identity. The system is built using combination of hierarchical clustering and decision tree. After removing noise, we achieved 9% average accuracy improvement over without removing noise and average accuracy of 73%.

Nature of decision tree is well suited for generating list of most favorable groups for user. In our future work, we will improve the GRS by listing a certain number of most suitable groups according to the users' profile. Tree figure on Fig. 1 illustrates that once the most suited group is found, other nodes in same sub–tree or neighbor share similarity with the most suited group. This property can be vital to find list of suitable groups.

The main concept behind the GRS can be used in many different applications. One is information distribution system based on profile features of users. As social networking community expands exponentially, it will become a challenge to distribute right information to a right person. We need to have a methodology to shape flooding information to user from his/her friends, groups, and network. If we know identity of the user's groups, we can ensure the user to receive information he/she prefers.

Another research area can be explored is targeted-advertising [4] to individuals on social network site. Many advertising technique are already implemented, such as Amazone based on users' search keywords and Google Adsense based on context around its advertising banner. In addition, Markov random field technique has emerged as useful tool to value network customer [5].

Acknowledgments. This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0619871and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would like to thank our anonymous reviewers for their insightful and helpful comments and suggestions.

References

1. Adamic, L.A., Buyukkokten, O., Adar, E.: A social network caught in the web. First Monday, vol. 8 (2003)
2. Backstrom, L., Huttenlocher, D.P., Kleinberg, J.M., Lan, X.: Group formation in large social networks: Membership, growth, and evolution. In: KDD, pp. 44–54 (2006)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Chapman & Hall, New York (1984)
4. Chickering, D.M., Heckerman, D.: A decision theoretic approach to targeted advertising. In: UAI, pp. 82–88 (2000)
5. Domingos, P., Richardson, M.: Mining the network value of customers. In: KDD, pp. 57–66 (2001)
6. Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.: Self-organization and identification of web communities. IEEE Computer 35(3), 66–71 (2002)
7. Flake, G.W., Tarjan, R.E., Tsioutsiouliklis, K.: Graph clustering and minimum cut trees. Internet Mathematics 1(4), 385–408 (2004)
8. Girvan, M., Newman, M.: Community structure in social and biological networks. PNAS 99(12), 7821–7826 (2002)
9. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. Journal of the American Statistical Association 97(460), 1090–1098 (2002)
10. Hopcroft, J.E., Khan, O., Kulis, B., Selman, B.: Natural communities in large linked networks. In: KDD, pp. 541–546 (2003)
11. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. PNAS 102(33), 11623–11628 (2005)
12. Romesburg, H.C.: Cluster analysis for researchers. Lulu Press, North Carolina (2004)
13. Viegas, F.B., Smith, M.A.: Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In: HICSS (2004)
14. Wellman, B., Boase, J., Chen, W.: The networked nature of cummunity. IT&Society 1(1), 151–165 (2002)