Finding Core Members in Virtual Communities*

Haiqiang Chen
Institute of Computing Technology
P.O. Box 2704
Beijing, China
chenhq@software.ict.ac.cn

Xueqi Cheng
Institute of Computing Technology
P.O. Box 2704
Beijing, China
cxq@ict.ac.cn

Yue Liu
Institute of Computing Technology
P.O. Box 2704
Beijing, China
liuyue@ict.ac.cn

ABSTRACT

Finding the core members of a virtual community is an important problem in community analysis. Here we presented an simulated annealing algorithm to solve this problem by optimizing the user interests concentration ratio in user groups. As an example, we test this algorithm on a virtual community site and evaluate its results using human "gold standard" method.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval; H.3.5 [**Information Systems**]: Online Information Services; J.4 [**Computer Applications**]: Social and Behaviorial Sciences

General Terms

Algorithms, Experimentation,Human Factors

Keywords

Virtual community, core members finding, simulated annealing

1. INTRODUCTION

People have been using virtual communities to communicate since the beginning of the Internet. In the last few years, many virtual communities burgeon on the Internet, and it is very easy now to build a new virtual community using tools provided by those web-based community building sites. In simple terms, virtual community (or online community) is the gathering of people, in online space where they can come, communicate, and get to know each other better over time. According to the definition of Howard Rheingold in [3], virtual communities are social aggregations that emerge from the Net when enough people carry on public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace.

An important feature of the virtual community is the open membership. One can join any virtual community he want, and he can reach all its members easily. The open access to virtual communities brings numerous members to them. But among those members, most are seldom heard to other members. Among those non-active members, some are pure observers, and some are light participants, raising voice occasionally. Only a small portion of the members are active in the virtual community. In our study, such active members are called *core members*.

Finding the core members in virtual community is an intriguing problem and has been little researched so far. In this paper, inspired by the ideas of Guimera[1] we proposed a computer algorithm to solve this problem. As an example, we give results from the application of this algorithm in the virtual communities of a Chinese web site Douban.com.

*This research is supported by the 973 National Basic Research Program of China (2004CB318109 & 2007CB311100) and MSRA IST 2007(FY07-RES-THEME-067).

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. THE ALGORITHM

According to our intuitive understanding and the definition by Rheingold[3], the shared interests and activities are one of the most important feature of the virtual community. Generally, the shared interests and activities are the major reason to attract users to join the virtual community.

In most of the virtual community web sites, the user tagging information can be used to define the interests of the users. We assume that one user u's interests can be presented by the set of subjects which have been tagged by this user. So user u has a set of tagged subjects: $I_u = \{s_1, s_2, \ldots, s_n\}$. As a result, we can get a bi-partite graph between all users and tagged subjects.

Considering user u and subject s: assuming that s has been tagged by M_s users, and user u has tagged N_u subjects, then, the probability that user u have tagged s is:

$$M_s \frac{N_u}{\sum_k N_k} \tag{1}$$

So, it can be easy to know the probability that s has been tagged by both u and v is:

$$M_s(M_s-1)\frac{N_u N_v}{(\sum_k N_k)^2} \tag{2}$$

The expectation of the number of subjects tagged by both u and v is (note that $\sum_s M_s = \sum_k N_k$):

$$\frac{\sum_s M_s(M_s-1)}{(\sum_s M_s)^2} N_u N_v \tag{3}$$

As $\sum_s M_s(M_s-1)$ and $\sum_s M_s$ are global properties which

do not depend on the pair of u and v selected, we can calculate equation (3) very quickly using some simple multiply operations. So given a user set $G = \{u_1, u_2, \ldots, u_n\}$, we can define the user interests concentration ratio as the cumulative deviation of the number of shared subjects tagged from the random expectation:

$$C_G = (\sum_{u \neq v \in G} c_{uv} - \frac{\sum_s M_s(M_s-1)}{(\sum_s M_s)^2} N_u N_v) / \sum_s M_s(M_s-1) \quad (4)$$

where c_{uv} is the number of shared interests tagged by both u and v. If the number of shared interests in group G is no different from the random expectation, then this quantity C_G will be zero. If the value C_G is larger than zero, it indicates that the users in group G have more shared interests than the random expectation.

In a virtual community, the core members will have more interaction with each other than those non-core members. As a result, the core members in a virtual community would share more common interests with each other than those non-core members. So we can use the user interest concentration ratio C_G to find the core members in virtual communities: the problem of finding the core members of a virtual community can be transformed into the problem of finding the portion of members in the virtual community with large C_G value.

In order to solve this problem, we choose to use the simulated annealing (SA)[2] method. Simulated annealing[2] is a stochastic optimization technique that enables one to find 'low cost' configuration without getting trapped by the 'high cost' local minima.

Our algorithm can now be defined as follows.

1. Set the initial temperature as T and initial solution $S = G$;
2. Randomly choose $a \in G$: if $a \in S$, then $S' = S \backslash \{a\} = \{x | x \in S, x \neq a\}$; if $a \notin S$, $S' = S \cup \{a\}$;
3. Calculate $\Delta = C_{S'} - C_S$: if $\Delta > 0$, accept the new solution $S = S'$, if $\Delta <= 0$, accept the new solution $S = S'$ with the probability $e^{\Delta/T}$;
4. Cool down the temperature $T = cT$, c=0.995;
5. Repeat step 2 - 4 until the temperature T is small enough.

3. RESULTS ON DOUBAN.COM

In order to test our algorithm, here we give one example, the analysis of the virtual communities in Douban.com. Douban.com is a Chinese web site. It can be labeled as a "collaborative filtering" or "collaborative tagging" site, where one can tag his interested items and share his reviews or comments about millions of books and movies with others. Basing on the user-tagging behaviors, reviews and comments, Douban.com helps its users to get recommendations about new books and movies.

As another result, Douban.com can also help one to find other users with similar tastes and interests, so they can get connected and communicate with each other. Douban.com provide a community service, which is called "Douban Group". Anyone can set up a new group about some topic and invite others to join this group. In these groups, group members can discuss their interested topic, and get more information from other members.

Score	Description
1	Full committed participant and group leaders.
0.5	The participant who is heading toward to be full committed, but still not now.
0	New comer or lurker who never talk in this group.

Table 1: User categorization table

	Human Rating				
Algorithm Result	0	0.25	0.5	0.75	1
0	0.39	0.02	0.005	0.01	0.055
1	0.24	0.04	0.02	0.02	0.21

Table 2: Correlation between human ratings and algorithm results.

In this study, we chose all tagged items by one user as his interests profile. Each item which has been tagged by a user is one of his interests. Our algorithm can output a list of core members for each group.

Since there was no explicit core members data in Douban.com, we needed to use human raters to generate "gold standard" to evaluate our results. As it was not possible for us to rate a large number of groups and users, we rate only a small set of the users. We randomly chose 10 middle-sized groups (group size varies from 100 to 200) from 38423 groups on Douban.com, and then for each group, 20 members were chosen as our evaluation samples.

We invited two raters who are also Douban.com users(but not the members of the 10 selected group). For each group, the raters explored the group's home page, bulletin boards and personal profiles of every members, so as to decide weather the 20 selected user was the core member of the group. Users were categorized as Table 1.

After both raters submit the rating results, we use the average score by the two rater as the final human rating results. Table 2 shows the correlation between human rating score and the algorithm output score for all of our 200 test samples. This results show that our algorithm can find most of the core members from the virtual community, although the false positive rate is quite high.

4. CONCLUSIONS

Finding the core members in virtual communities is a intriguing problem for social network analysis and other related areas. Here we presented a SA algorithm to solve this problem. We also apply and evaluate this algorithm in a real-world online community, Douban.com. The algorithm show some satisfying results. But how to improve this algorithm's false positive rate is still a major challenge for us.

5. REFERENCES

[1] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76, 036102(2007).

[2] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[3] H. Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier, revised edition.* The MIT Press, November 2000.