

## *A Study on Finding Potential Group of Patrons from Library's Loan Records*

Toshiro Minami<sup>\*\*†</sup> and Kensuke Baba<sup>†</sup>

<sup>\*</sup>Kyushu Institute of Information Sciences, Japan

<sup>†</sup>Kyushu University Library, Japan

### Abstract

Social networking services that connect a person to other people are attracting people's attention and various types of services are provided on the Internet. Library has been playing a role of social media by providing us with materials such as books and magazines, and with a place for reading, studying, getting lectures, etc. In this paper, we present a method for finding candidates of groups of the library's patrons who share interest areas by utilizing the loan records, which are obtainable in every library. Such a homogeneous group can become a candidate for a study group, a community for exchange ideas, and other activity group. We apply the method to a collection of loan records of a university library, find some problems to be solved, and propose measures for more detailed solutions. Even though the potential group finding problem still remains a lot of issues to be solved, its potential importance is very high and thus to be studied even more for future applications.

**Key words :** Library Data Analysis, Patron/Student Profiling, Interest Area, Social Network Discovery

### 1. INTRODUCTION

Due to the recent development of ICT (Information and Communications Technology), quite a lot of people rely on the Internet services today. Especially social networking services attract people's attention. Library has been providing with social services and plays a role like a social media. Considering the vigorous change of information access environment, libraries have to renovate their service model. In this paper, we present a method for finding candidates of groups of the library's patrons who share interest areas. Such a homogeneous group can become a candidate for study group (see [5] also), a community for exchange ideas, and other activity group. The interest area data as a patron profile is defined from the library's loan records as in the same way as our previous studies [1-4]. Interest areas are useful in order to capture the features of a patron and to use them for supporting the patrons with learning new knowledge. For example, university libraries

are supposed to help their student patrons with learning in the ways not only with just providing learning materials and study rooms but also with advising them what to learn, how to learn, with whom to learn, etc. based on the patron's learning history, learning style, and other features in learning. Such services are very important for a university/academic library as a subordinate organization of a university, where its major mission is to educate students.

Based on such a motivation, we deal with the method of finding homogeneous groups of patrons using the similarities of the profiles of student patrons which are obtained from the loan records of university library. Such methods are considered to be applicable also to other types of libraries such as public libraries because the library's main mission can be considered to educate their patrons in their information skills as well as to providing with study opportunities.

### 2. LOAN DATA AND FUNDAMENTAL CONCEPTS FOR ANALYSIS

The library's loan records we use in this paper are provided by the Central Library of Kyushu University, Japan, for the academic year 2007 (from April 2007 to

March 2008) [1-3]. A record consists of the book ID, book's NDC (Nippon, or Japan, Decimal Classification) number, call number, borrower's patron ID (renumbered one by considering the privacy issue of the patron), affiliation, and type, and the timestamps for borrowing and returning dates and times, etc. After the preprocessing by eliminating the records that lack necessary information, 53,182 records are left, which becomes the target data for analysis. As many as 5,718 patrons are remained after preprocessing.

One of our aims of this paper is to find a (homogeneous) group of patrons who have similar, or common, interest areas. To begin with, we define the concept of profile for interest areas using the LR (Loan Records) [1-4]. We choose the NDC numbers for categorizing the "areas" and define the "interest areas" of the patron by the ratios of the numbers of borrowed books of the NDC numbers by the patron. NDC is a hierarchical system where the top-most level consists of 10 categories from 000 to 900, and each category is divided into another 10 subcategories, and such sub-categorization process continues several times. The top level categories of NDC are; 000 (General Works), 100 (Philosophy and Religion), 200 (History and Geography), 300 (Social Sciences), 400 (Natural Sciences), 500 (Technology or Engineering), 600 (Industry and Commerce), 700 (Arts), 800 (Language), and finally 900 (Literature).

We define the interest area profile as a 10 dimensional vector by the element corresponds to the top-level category of NDC. The value of each element of the vector is the ratio of the number of borrowed books from the corresponding top category of NDC. It is define formally as follows: Let  $p$  be a patron. The profile of  $p$   $\text{Prof}(p) = \langle r_{000}, r_{100}, \dots, r_{900} \rangle$ , where  $r_i = f_i / \sum_{j=000}^{900} f_j$  and  $f_i = \#\{l: a LR | \text{borrower of } l = p, \text{ category of } l = i\}$  for  $i = 000, 100, \dots, 900$ . We take the cosine similarity as the similarity measure for the profiles.

### 3. FINDING THE CANDIDATES OF HOMOGENEOUS GROUP OF PATRONS

In this section, we investigate and develop a method of finding candidate group of patrons with similar interest areas by using the loan records described in Section 2. We calculate the similarities of all the 16,344,903 pairs of students. Then we can see that most pairs (9,072,290, 55.5%) have similarity of 0. Further 1,383,927 pairs (8.5%) have similarity of 1, which means that the interest area vectors  $A$  and  $B$  have relations  $A = kB$  for some non-zero number  $k$ . This relation of similarity of 1 gives an equivalence relation and we can partition the students into groups using this relation.

Fig. 1 shows the numbers of groups with the sizes marked

in the x-axis. The sizes vary from 1 as minimum and 1,383 as maximum. From the figure we can see that there are a lot of small groups. For example, 1,617 groups are of size 1; i.e. the patron who have no other patrons with exactly the equal interest area pattern. Also we can see there are 114 pair groups and 42 groups with three members having the equal interest area pattern.

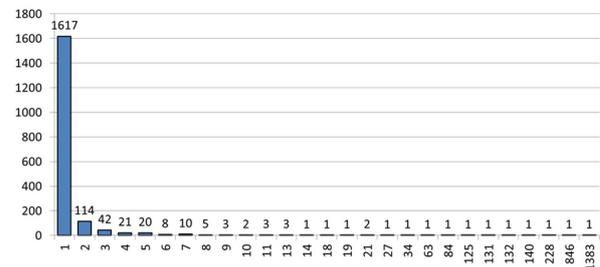


Figure 1. Numbers of Groups According to the Group Size

We discuss this grouping method from the view of taking it for study together. Size is important for study groups. If a group size is too big, i.e. it consists of quite a lot of members, it is very hard to let all the members study appropriately. On the other hand a group with small size may cause a trouble. For example, such cases may happen often that one member has to teach other members always, or noone can understand what is written on the text book.

Even though the most appropriate number of members depend of situations such as the matching of the members, what subject or textbook to learn together, the way how to run the study group etc. From our experience, the most appropriate group size might be 4 to 6 in a typical study group.

According to Fig. 1, the 64 groups having the group size from 4 to 8, i.e. 342 patrons in total, can be recommended as study groups as they are. Let us choose a group which consists of 5 members. Even though the numbers of books borrowed by them vary from 27 to 9, their profiles are the same in the sense they borrowed the books from NDC 400 (Natural Sciences) and 500 (Technology and Engineering) in the ratio of 8 to 1. For the groups with the size from 1 to 3 (i.e. for 1,971 patrons), it is recommendable to extend the group size by adding 1 to 4 or 5 members who have high similarity to the members of the groups. It is more difficult for the groups with the size greater than 8; i.e. from 9 to 1,383. Even though the group size is too large, the members profiles are exactly the same (except the number of borrowed books), we can not decide how to divide them into subgroups with appropriate sizes.

In order to develop an appropriate method for dividing the large groups, we need to find a new measure for further investigation. One idea for new criterion is to introduce more sophisticated similarity measure for more precise

comparison of interest areas. We take a closer look at the largest group, which consists of 1,383 members. Even though the numbers of borrowed books of this group vary in a wide range from 127 to 1, all of them borrowed only from the NDC 400 (Natural Sciences). The second largest group consists of 846 members, and the numbers of books borrowed by one patron vary from 63 to 1. All the members borrowed with NDC 300 (Social Sciences). As we inspect other groups, we find that other big size groups also consist of patrons with very limited interest areas. For example, the third largest group is for NDC 500 (Technology and Engineering), the fourth largest one is for NDC 900 (Literature), and so on.

As we see the ratios of the number of patrons according to the numbers of NDC categories (from 1 to 10) of a patron, nearly 60% of patrons borrow the books only from 1 NDC category. Further, about 1/4 of patrons borrow books from 2 NDC categories and 1/9 of patrons from 3. Thus about 90% of patrons borrow books only from 1 to 3 NDC categories. As we see the profiles of top 11 students according to the numbers of borrowed books, 2 patrons borrowed books from 3 NDC categories and the other 9 borrowed from more than 3 categories, and 2 patrons borrowed from all 10 categories. These students are rather exceptional.

Now we investigate deeper about the group with maximum size where patrons borrowed only the books of natural sciences (NDC 400). We take the similar approach to the subcategory of NDC 400; i.e. from 400 to 490. The total number of books borrowed by them is 7,682. The books in the category 410 (Mathematics) are borrowed most among these 10 subcategories. The next largest is 430 (Chemistry) followed by 420 (Physics), 460 (Biology), 450 (Earth Sciences), and so on.

As in the same way to top level NDC categories we calculate the similarities of 955,653 pairs of 1,383 patrons. As we see the frequencies for 0, 1, and between 0.0 and 1.0. This time again, quite a lot (657,799, 69%) of pairs have similarity value 0. Among the rest classes similarity value 1 takes the most frequency (67,293, 7%). In the similar method as above we classify the patrons with the relation of similarity 1. Then the biggest group consists of 205 patrons, who borrowed only books of NDC 420 (Mathematics). As the result of using the subcategory of NDC with the order of the number 10, the group size becomes very small in comparison with using only the top level NDC numbers of the order of 100. However the biggest group size of 205 is still very big. We still need to put even more measures to divide such groups.

The first issue is about the fundamental questions on our approach: "What students are good as a member of study group?" and "How can we find the possibly good group from the loan records (and other data)?" These questions are

very fundamental and difficult to answer. As the first step to answer these questions, we try to find some necessary conditions. At least we can say if the interest areas of the group members are separated vigorously, the group will have some difficulty in what to learn together. So in this paper, we assume first of all that a study group should be formed so that the members' interest areas are somewhat in common. We would like to assume also that the knowledge levels of the members should be somewhat close enough so that to avoid the case that a specific member who has very high knowledge level always teaches other members who are in the low level. We have already studies about the expertise level estimation [1], the result of which could be applicable in this respect.

It is necessary to discuss about the size, i.e. the number of members, of the group. If a group size is too large it is quite difficult to keep the group work effectively. From our experiences, the most appropriate size might be from 4 to 6 or something in a normal situation. It might vary according to the situation. We need more investigation on this issue.

In order to form a group in the suitable size, we need to introduce other criteria or measures so that to make the decision method more precise. One possibility is to set a threshold value for decision making. For example, suppose we set a threshold value, say 0.9, for similarity. Then we consider two patrons having similarity 0.9 or greater as in the same group. So we can make a bigger group in this way as we compare with the case of using the value 1 only for making a group. We can adjust this method of grouping by changing the threshold value.

Another way of tuning the method is of using the modified profiles of interest area. In the previous discussions we only consider the plain cosine similarity as the measure for similarity. Suppose a patron happens to borrow a book in an NDC category he or she has never borrowed before. This category is not so important for the patron. However, if another patron borrowed quite a lot of books in this category, the resulting value of inner product of these two patrons' profiles becomes very big as if their similarity of interest area is very high. Such a case may happen.

In order to avoid such cases, we can modify the original definition of profile as follows: For example, we set the profile value only when the patron borrowed books in this category with bigger ratio than the pre-set value, say 0.1, and set the profile value to 0 otherwise. Then only those categories the patron really interested in have non-zero values and thus effect to the calculation of similarity.

In our first thought, clustering techniques can be well used for study group formation. However it is not always appropriate. If a student wants to learn several subjects, the study groups could not set in the same time zone, which

does not match to the concept of clustering. On the other hand, clustering methods are useful if a number of students want to study in small groups in the same time zone. Some kinds of lectures prefer to take this type of study style.

In more general situation, one student can join more than one study groups and he or she studies some number of subjects in the several study groups. In such a situation, the study group finding method might be supposed to find several candidates that seem to be appropriate for studying different subjects. We have to investigate the methods for such situations in our future work.

#### 4. CONCLUDING REMARKS

As a part of studies for library marketing, we discussed the method of forming homogeneous groups of patrons in this paper. We put special focus on students' study group in a university. We dealt with the interest area profiles of students and recommend a group of students as the candidates of study group based on the similarity of students. We applied this method to a collection of loan records from Kyushu University Library and concluded that the study group recommendation could be very useful.

We also discussed what we can do in order to adjust our method of homogeneous group formation. By setting a threshold in the process of deciding if a patron and another patron are appropriate to be in the same group or not, it will become easier to form a group in an appropriate size. Another idea is to modify the profile vector so that non-zero values are assigned to the only NDC categories that the patron is interested in. Then the subject areas the patron is not so much interested in will not give effects to similarity values; they are ignored in group formation as well.

We have to keep on refining our method in this direction as well as to attack the issues: (1) to search for more possible options in each step of our research scenario and to investigate more precisely to clarify which option is most appropriate in each situation, (2) to add up other types of information to the currently considering factors of interest area, (3) to extend our model so that our method can recommend appropriate textbook(s) for the recommended study groups, and, last but not least, (4) to experiment with the actual groups of students studying in the recommended groups and feed-back the results.

#### ACKNOWLEDGMENT

This research was partly supported by the Ministry of Education, Science, Sports and Culture, Grant-in- Aid for Scientific Research (C), 24500318, 2013.

#### REFERENCES

- [1] Minami, T.: Expertise Level Estimation of Library Books by Patron-Book Heterogeneous Information Network Analysis -- Concept and Applications to Library's Learning Assistant Service --. The 8th International Symposium on Frontiers of Information Systems and Network Applications (FINA 2012), DOI 19.1109/WAINA.2012.184, pp. 357-362 (2012).
- [2] Minami, T.: Interest Area Analysis of Person and Group Using Library's Circulation Records, Proc. IADIS International Conference Information Systems (IS 2013), 8pp. (2013).
- [3] Minami, T., Baba, K.: An Attempt to Find Potential Group of Patrons from Library's Loan Records, International Symposium on Advanced and Applied Convergence (ISAAC 2013), 4pp., 2013.
- [4] Minami, T., Baba, K.: Library Patron's Interest Area Profiling and Similarity Based Grouping --With an Investigation on Applicability to Study Group Formation --, International Conference on Applied and Theoretical Information Systems Research (ATISR 2013), 12pp. (2013).
- [5] O'Hearn, K.: Computer-Aided Student Study Group Formation, McNair Scholars Journal, vol. 13, iss. 1, Article 10 (2009)  
<http://scholarworks.gvsu.edu/mcnair/vol13/iss1/10>.



**Toshiro Minami** received his BS degree from Kyushu Institute of Technology in 1973, and MS and D.Sc. degrees from Kyushu University in 1975, and 1999, respectively. He was a researcher of Fujitsu Limited and Fujitsu Laboratories Limited from 1984 to 1999, a research fellow of Australian National University from 1992 to 1993, an associate professor of Kyushu University Library from 1999 to 2001. He has been a professor of Kyushu Institute of Information Sciences since 2001. His research interests include multi-agent system, data mining and library informatics.



**Kensuke Baba** received his BS, MS, and D.Sc. degrees from Kyushu University in 1996, 1998, and 2002, respectively. From 2002 to 2003 he was a research fellow and from 2003 to 2009 he was an assistant professor in Faculty of Information Science and Electrical Engineering in Kyushu University. Currently, he is an associate professor in the library of Kyushu University. He is a member of IEEE. His research interests include data mining, pattern recognition, and digital library.