IADIS International Journal on Computer Science and Information Systems
Vol. 8, No. 1, pp. 112-129
ISSN: 1646-3692

AN ANALYSIS OF INTEREST AREA SIMILARITIES BY UTILIZING THE LOAN RECORDS OF LIBRARY

Toshiro Minami. *Kyushu Institute of Information Sciences.6-3-1 Saifu, Dazaifu, Fukuoka 818-0117 Japan.*

ABSTRACT

The first aim of this paper is to make a modeling framework of the interest area of a library patron by utilizing library's loan records. In other words, to profile a person with his/her interest field. Interest area profiling is not only interesting of itself but also useful for library when it chooses the books to be purchased, for professors when they give lectures, and for other people. The eventual goal of the study presented in this paper is to develop various data analysis methods that provide with useful tips in assisting people, including library patrons, as they learn. The second aim is to develop some useful analysis tools for such pedagogical purpose, especially for analyzing the library data and lecture data. In this paper we propose a concept of virtual faculty of a member of university, which is defined based on the interest area profile of a patron (student) and a faculty. We also give a comparative study between patron students and faculties. Even though the approach and analysis methods presented in this paper are rather in the primitive stages of the research toward this direction, they have a high potential and are expected to be developed to be matured and be practical in the near future.

KEYWORDS

Knowledge Management, Knowledge Discovery, Library Marketing, Data Analysis, Data Mining, Library Data

1. INTRODUCTION

Thanks to the development and popularization of ICT (Information and Communications Technology) our society is getting to be knowledge-oriented and it will be going on toward the future. Libraries have been playing an important role in society as a public service that provides us with a strong support when we study and acquire knowledge. They are supposed to keep playing even more important social role in knowledge and skill acquisition in the knowledge-oriented society. Especially, university libraries are supposed to strengthen their learning supporting services even more because they belong to educational organizations of universities.

In order to obtain effective plans for strengthening learning assisting functions, it is a hopeful approach to develop data analysis methods from various data sources such as library data, lecture data, and other ones. By using such objective data, the tips and know-hows obtained in the data analysis are supposed to be also objective. In this paper, we use loan records from Kyushu University Library in Japan. We have pursued a couple of case studies and have shown the usefulness of library's loan records. We define the concept of interest area as a profile of a library patron and show what kinds of characteristic features of a patron we can see from the profile [8]. We also define the profile of a group of patrons in a similar way. We apply this concept to the faculties of Kyushu University and see some characteristic features of faculties.

We define two user-profiling measures so that we can compare user-profiles easier than just see and check the profiles themselves. We compare a group of patrons and the faculties using these measuring indexes. We then define the similarity measure of two profiles using cosine similarity. With this measure we can investigate the similarities between patrons, between faculties, and patron and faculty. We can define the concept of virtual affiliation, or faculty, of a patron as the faculty that has the highest similarity in its interest area profile in comparison with other faculties. We can see that many patrons have the same virtual faculty as the real one the patron belongs to, but some patrons have different virtual faculties instead of the real ones.

The rest of this paper is organized as follows: In Section 2, we describe the target data for analysis. In Section 3, we define the concept of interest area profile of a patron, and define some concepts for comparing the features of students' and faculties' interest areas. We also conduct the comparative study of the relations between patrons and faculties. Based on these studies we define the concept of virtual faculties in Section 4, and discuss what we can find from the results. And finally in Section 5, we summarize what we have done in this paper and prospect our future works.

2. TARGET DATA FOR INTEREST AREA PROFILING

We use the loan records obtained from the Central Library of Kyushu University, for the academic year 2007; i.e. from April 2007 to March 2008, which were used also in our papers [4-6, 8-9]. The data contain 67,304 loan records. A record consists of the book ID, book's NDC classification number (Nippon Decimal Classification), call number, borrower's patron ID (renumbered in considering privacy), borrower's affiliation, borrower's type, and the timestamps for borrowing and returning dates and times, etc. The number of patrons who borrowed at least one book is 6,118 and the average number of books per patron is about 11.

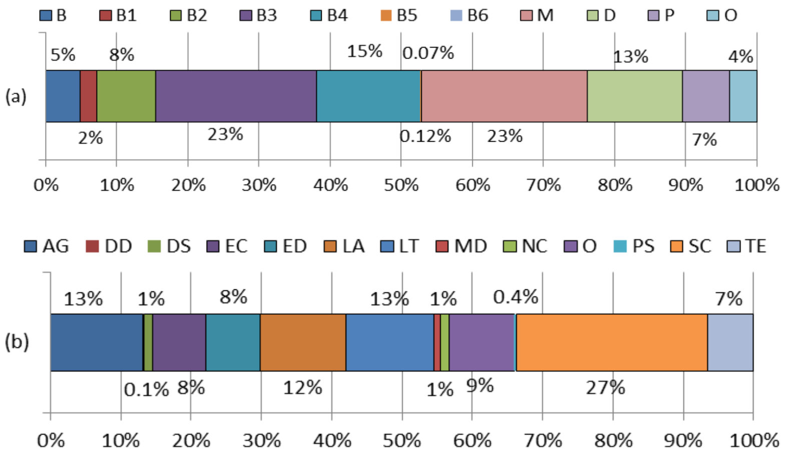

Figure 1. The Borrower Ratios in Percentage According to: (a) Patron Types, and (b) Affiliated Faculties

Figure 1 shows the ratios of patron types and affiliations. The borrower types are divided into 10 types; undergraduate (Bachelors-1 to 6, or B1 to B6), masters' (M), Ph.D (D), academic staff (Professors, P), and others (O). The faculty names stand for (in lexicographic order), AG for Agriculture, DD for Dental, DS for Design, EC for Economy, ED for Education, LA for Law, LT for Letter, MD for Medicine, NC for the special faculty called 21st century program, which is for the students who wish to study a wide variety of fields, O for whom that do not belong to other faculties, PS for Pharmaceutical, SC for Sciences, and TE for Engineering. It is easy to see in Figure 1 that students are the majority borrowers and B3 and M students occupy the big shares among students. Also we can see that SC (Sciences) overcomes other faculties, followed by AG (Agriculture) and LT (Letter).

In the preprocessing of the data, we eliminate the records that have inappropriate values and even have no values for the inevitable properties (items) that are necessary to deal with in the analysis presented in this paper. For example 244 records have NDC numbers that are greater than 1,000 and 7,260 records have the non-numeric values for this item and thus have eliminated from the original records. There are 53,182 records that are left after eliminating such records. The number of patrons in these remaining records is 5,718. As the result, about half (53%) of books are borrowed by undergraduate students and 23% by masters' and 13% by Ph.D students. Thus about 89% of books are found to be borrowed by students.

3. INTEREST AREA PROFILE AND COMPARATIVE STUDY OF SIMILARITIES BETWEEN STUDENTS AND FACULTIES

This section deals with interest area profile [8, 9]. First in Section 3.1 we define the concept of profile by using the loan records and show the profiles of some representing patrons and of faculties. In Section 3.2, we define interest range and strength of an interest area profile. We also define the concept of similarity of profiles as a preparation for investigating characterization of the profile of student in a comparison to faculty's interest area. In Section 3.3, we try to optimize the order of faculties for better visualization in comparison to students' interest areas against those of faculties. Finally in Section 3.4, we conduct the actual comparisons of students' profiles against those of faculties.

3.1 Interest area of a Patron and a Group of Patrons for Profiling

The intending aim of defining the concept of interest area is to understand the patrons in terms of their eagerness, style, preliminary knowledge, etc., for learning. Our interest in this paper is on analyzing the areas of interest of a patron, or what subjects or topics the patron is interested in. We are hoping to obtain information about the patron's expertise field together with the related field he/she wishes to learn.

The concept of the profile of a patron in this paper is defined by using the library's loan records [8, 9]. For the areas of topics, we use the NDC number which is assigned to the books as a part of their bibliographical information. NDC is a decimal classification system like the DDC (Dewey Decimal Classification) system localized to Japan. The top level categories of NDC consist of the following 10 topic fields; 000 for General Works, 100 for Philosophy and Religion, 200 for History and Geography, 300 for Social Sciences, 400 for Natural Sciences, 500 for Technology (Engineering), 600 for Industry and Commerce, 700 for Arts, 800 for Language, and 900 for Literature. Note that NDC is different from DDC.

We define the profile of a patron as a vector with dimension 10, with each element corresponds to one of the 10 top categories of NDC. An element of the vector is the frequency of the borrowed books of the patron which have the corresponding top category numbers of NDC. Thus, for example, if a patron borrows 11 books with NDC number from 100 to 199.99, 12 books from 200 to 299.99, and so on until 19 books from 900 to 999.99, the profile vector of the patron becomes <11, 12, 13, 14, 15, 16, 17, 18, 19>. We can extend this definition to a group of patrons by just modifying the condition from "borrowed by the patron" to "borrowed by one of the patrons of the group." It is possible to define other concepts in a similar way such as the profiles of properties of patron that relate more on knowledge level, learning ability, learning style, etc.

Figure 2 shows the interest area profiles of the top 11 patrons according to the number of items, or books in the left graph, and the interest area profiles of the faculties in the right graph. We chose them because firstly they are representative patrons among all the patrons and knowing them is important for library marketing, and secondly because quite a lot of patrons borrow only a couple of books and thus they are not appropriate to use as sample data for developing new methods for profiling the patrons. The top-most patrons from A to K (also called by P.A to P.K) borrow as many as between 388 and 143 books during one year.

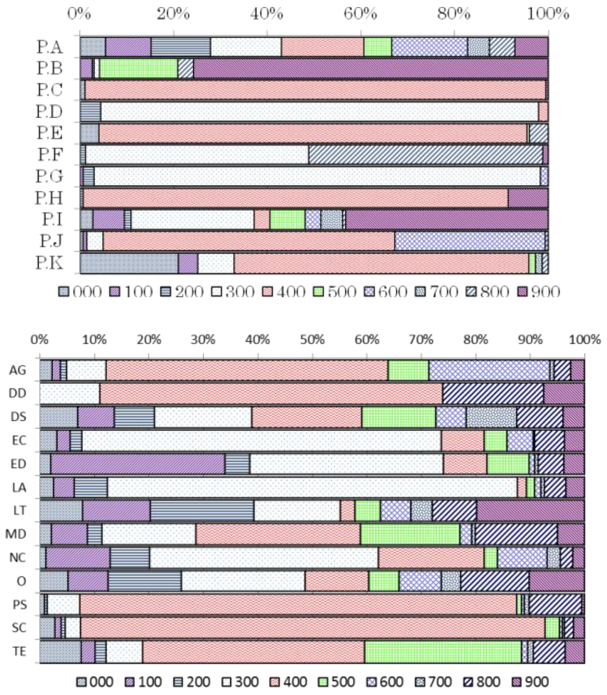

Figure 2. (left) Interest area Profiles of the Top 11 Patrons According to the Numbers of Borrowed Books, or Items, and (right) Interest area Profiles of Patrons' Affiliations, or "Faculties"

It is easy to see in the figure that the ratios of books according to the classification number, or topic area, vary from patron to patron. For example, P.A borrows quite a wide area of books with NDC numbers from 000 to 900. On the other hand, P.C borrows mostly with the classification number 400 (Natural Science). Such difference of the width of topic areas indicates a character of the patron in his or her interest range, or curiosity range. Together with the number of the borrowed books, this range can be good measures for characterizing features of a patron, which will be discussed in more detail in the next section.

From the right part of Figure 2, we can see that the faculties PS (Pharmaceutical) and SC (Sciences) have a very high top interest area at the NDC number 400 (Natural Sciences) and they are similar in this respect. On the other hand the faculties DS (Design), LT (Letter), and O (Other) have relatively low value in the top interest area and they have a wide range of interest areas. These results are somewhat matching to our intuitive images on these faculties. In this respect, it is interesting to see that DD (Dental) and MD (Medicine) have relatively wide interest areas, which is against our naïve intuition.

3.2 Concepts of Interest Area's Strength, Range, and Similarity

We start this section with defining two measures of an interest area profile. They are intending to represent some characteristic features of the profile in two different points of view. The first one is the strength or magnitude of the interest of the patron and another is the width of the areas that the patron is interested in. Using such measures it is easier to compare the profiles. According to our previous work in [8], we survey the results on these concepts.

We define the concept of interest strength by the number of books, or items, that are borrowed by the patron or the group of the patrons. We take this definition based on the thought that if a patron with high curiosity would borrowes a lot of books and thus it might be a good measure for the strength of interest.

We define the interest area range, or range size, by the information entropy of the profile by using the ratios of the 10 NDC categories. Let pi = number of the books that belong to the NDC category i divided by the total number of the books, or the strength, of the patron's profile. Then the information entropy of the profile is calculated as the sum of -pi log pi. We use 10 for the base of the logarithmic function in order to make the maximum value to 1 because there are 10 NDC categories.

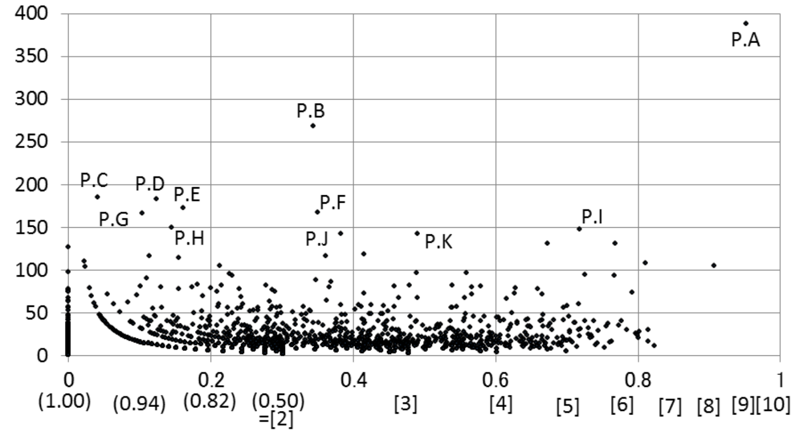

Figure 3. Correlation between the Range (x-axis) and the Strength (y-axis) of all the Student Patrons

Figure 3 shows the correlation between the range size and the strength of all patrons and Figure 4 shows those of the faculties. The range value 0 means that the patron borrows only one book. The range value is 1 if the patron borrows the books with all the NDC numbers, i.e. from 000 to 900, exactly the same number.

The patrons from P.A to P.K are named according to the order of the strength, or the number of borrowed books, so they are located in the upper part of the left graph. As has been predicted the range of P.A (0.952) is quite high; the highest among all patrons, so that it is located to the right-most and top-most place, which means he or she borrows the books from all the NDC categories with borrowing almost the same number of books each. Furthermore P.A borrows nearly 400 books, which are over 100 books more than the second patron, i.e. P.B, who borrows more than 250 books. P.A belongs to the other group (O) so that this patron represents the high interest range of the O group.

On the other hand P.C has the minimum range value (0.04), whose affiliation is SC and the year 4 undergraduate student (B4). This case also, the patron P.C is representing the characteristic feature of the faculty SC of having low interest range size. Like P.C the patrons P.D, P.E, P.G, and P.H are located in the left most part of the left graph with having the values less than 0.2, which means they borrow books with one category more than 80% of times and other ones less than 20%. Thus they have very limited range of interest. The patrons P.B, P.F, P.J, and P.K are located in the range with the range value from 0.3 to 0.5, which means, in roughly speaking, they mainly borrow books with 2 or 3 categories.

Among the best 11 patrons who are marked from P.A to P.K, there are 4 students with affiliation of SC (Sciences) in all, and 2 of them are B4, undergraduate at year 4, (P.C and P.H) and 1 (P.E) is B3 and another one (P.K) is M (Masters). The 3 undergraduate students have very low range values from 0.04 to 0.16. They are very concentrated in learning as the representing patron, and student, P.C. It is interesting to see that the remaining master's student (P,K) has relatively bigger range value 0.49. He or she borrows the books not only in the natural science field (with NDC 400), but also the books in general topics (with NDC 000), social sciences (with NDC 300) and others as well.

There are 3 Ph.D students with affiliation LA (Law); namely P.D, P.F, and P.G. The patrons P.D and P.G have similar range values 0.12 and 0.10, whereas P.F has bigger value 0.35 than the two. The former 2 students borrow the books with NDC 300 (Social Sciences) mostly, whereas the latter student borrows not only the books of social sciences but also the books with NDC 800 (Language) as many as of 300.

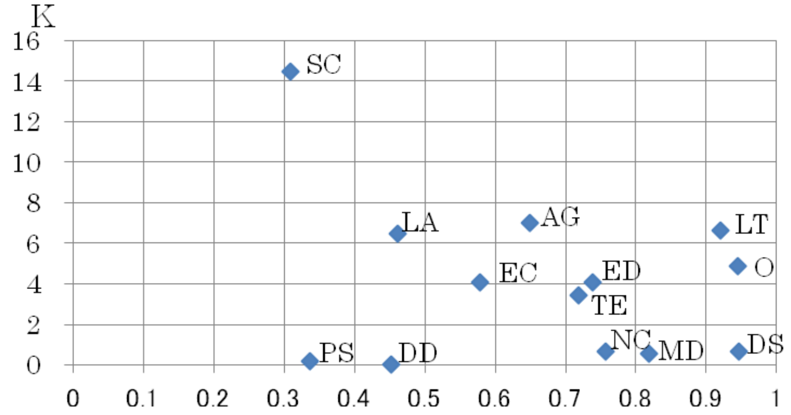

Figure 4. Correlation between the Range (x-axis) and the Strength (y-axis) of the Faculties

As we observe Figure 4, we can see that SC is far away from other faculties in both axes. It has the lowest value in region size and the highest in strength. Together with the right part of Figure 2, we can see that patrons in SC borrow the books in natural sciences (NDC 400) mostly and the number of the borrowed books are quite high, which probably means that their places physically locate very close to the library and it is quite easy for them to visit the central library and borrow a lot of books.

The faculties of PS (Pharmaceutical), DD (Dental), and LA (Law) are located in the left part from the line with the range size 0.5, which means that their patrons also borrows books of their expertise area mainly than other faculties. The reason why the numbers of borrowed books of PS and DD might be that their faculties locate in a different campus from where the library is, and the patrons in PS and DD visit the library in order to get the books they could not find in the libraries of their own campus. LA is, on the other hand, located in the same campus as the library and also the number of the members is larger than that of PS and DD.

It is interesting to see that DS (Design) and MD (Medical) are located in the lower right part of the graph where their range size is relatively large. Even though MD locates in the same campus as PS and DD, its range size is far bigger than these two. In order to find the reason of this fact, we investigate more on the patrons' behavior. Anyway in some reason the members of MD visit the library in a different campus in order not to find the books relating to their study in their expertise field but to find books in a wide variety of books.

DS locates in a campus of it own, i.e. different campus from that of library and even farther than the campus for MD, PS, and DD. The strength, i.e. the number of borrowed books, is small probably because of this reason. DS is a faculty that relates both to engineering and design, and thus it is easy to guess that their interest range as a whole is wide. However it is still surprising that its range size is larger than any other faculties including O (Other, or unclassified) and that LT (Letter) also has high range size. The members of LT borrow not just the books of literature (NDC 900), but also those in other areas as many as of literature.

Even if we can find similarities between two profiles just by looking them, it is often difficult to say how much similar they are. For example if we have three profiles, say A, B, and C, and we can “feel” that these 3 are similar, it is often very difficult to judge which is more similar between A and B, and A and C. In order to make such comparisons easier we introduce a new similarity measure between two profiles.

Since a profile is a 10-dimensional vector, we can use the cosine similarity. We define the similarity of 2 profiles P and Q as follows. Let P=<p1, p2, …, p10> and Q=<q1, q2, …, q10>. Then we define: Sim(P,Q) = P.Q / (||P||.||Q||), where P.Q is the inner product of P and Q; sum of pi.qi for all i from 1 to 10, and ||P|| and ||Q|| are the length of the vectors P and Q, respectively, where the length of P is defined as the sqrt(sum of pi.pi for all i from 1 to 10). The similarity value ranges from 0 to 1 as the value is the cosine of the angle between 2 vectors in the 10-dimensional Euclidean Space with the non-negative component values and thus the angle between them is from 0 to 90 degrees.

3.3 Similarity Analysis of Interest Area Profiles of Top 11 Students in Comparison with Profiles of Faculties

In this sectiion, we would characterize the interest areas of a student in comparison with the similarities of interest area profile against those of faculties. Figure 5 shows the similarities of the 11 patrons from P.A to P.K with the 13 faculties; 12 ordinary faculties plus 1 for other affiliations. We can see the similarity values make 2 clusters with high similarity and low similarity for the faculties AG (Agriculture) and TE (Engineering) clearly. The values for DD (Dental), EC (Economy), ED (Education), LA (Law), PS (Pharmaceutical), and SC (Sciences) also have two clusters, where the values may be expressed differently with two clusters plus one exception. The faculties like MD (Medicine), NC (New Century, or 21st Century), and O (Other) have no clear clusters.

As we have closer look at the ups and downs of the similarity values of patrons, we recognize that some faculties are close each other. For example, the values for SC, PS, and AG are somewhat similar for every student. Also the values for EC and LA are somewhat in common. If the similarity value to EC is high then that to LA is also high, and the value to EC is low then the one to LA is also low.

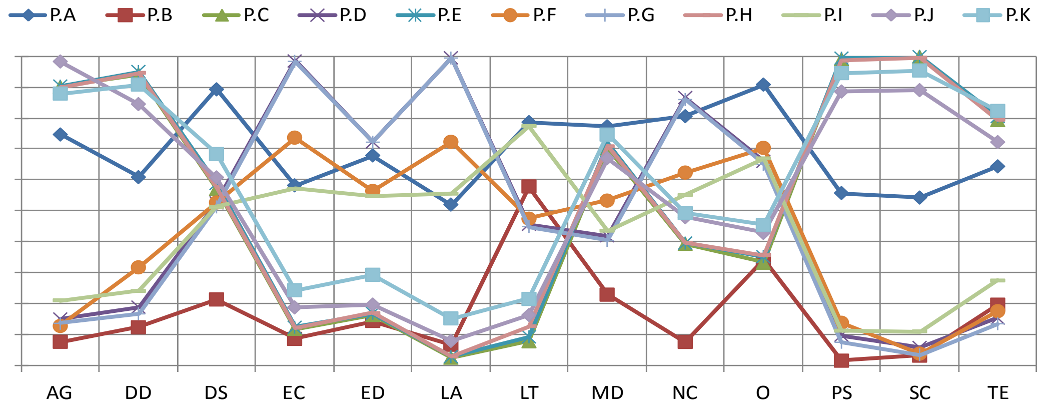

Figure 5. Similarities of 11 Patrons in Comparison with 13 Faculties

Considering this observation, it must be more convenient to optimize the order of the faculties so that the faculties with resemble similarity values are located close to each other and the similarity vallues for the student does not change vigorously.

In order to find the appropriate order of the faculties, we start with calculating the similarity of profiles for every pair of faculties in the same way for patron-and-patron, and patron-and-faculty. Table 1 shows the results.

Table 1. Similarities of 13 Faculties (in terms of the cosine similarity of profiles)

	AG	DD	DS	EC	ED	LA	LT	MD	NC	O	PS	SC	TE
AG	1	0.89	0.72	0.28	0.29	0.16	0.26	0.79	0.55	0.53	0.91	0.91	0.83
DD	0.89	1	0.7	0.3	0.3	0.21	0.26	0.83	0.53	0.54	0.98	0.95	0.8
DS	0.72	0.7	1	0.64	0.7	0.57	0.68	0.92	0.81	0.89	0.64	0.61	0.82
EC	0.28	0.3	0.64	1	0.77	0.99	0.54	0.55	0.92	0.76	0.2	0.16	0.28
ED	0.29	0.3	0.7	0.77	1	0.77	0.68	0.61	0.86	0.76	0.23	0.2	0.37
LA	0.16	0.21	0.57	0.99	0.77	1	0.54	0.46	0.89	0.72	0.1	0.06	0.18
LT	0.26	0.26	0.68	0.54	0.68	0.54	1	0.52	0.63	0.88	0.14	0.12	0.32
MD	0.79	0.83	0.92	0.55	0.61	0.46	0.52	1	0.71	0.78	0.77	0.74	0.91
NC	0.55	0.53	0.81	0.92	0.86	0.89	0.63	0.71	1	0.86	0.46	0.43	0.48
O	0.53	0.54	0.89	0.76	0.76	0.72	0.88	0.78	0.86	1	0.42	0.38	0.54
PS	0.91	0.98	0.64	0.2	0.23	0.1	0.14	0.77	0.46	0.42	1	0.99	0.82
SC	0.91	0.95	0.61	0.16	0.2	0.06	0.12	0.74	0.43	0.38	0.99	1	0.82
TE	0.83	0.8	0.82	0.28	0.37	0.18	0.32	0.91	0.48	0.54	0.82	0.82	1

From Table 1, we can see that the similarities between LA (Law) and SC (Sciences) has the minimum value of 0.06; which indicates that these two faculties are mostly different in their profiles, thus they would represent the two of the most extreme cases. So we choose the values of similarities to LA for x-axis and those to SC for y-axis. Figure 6 shows the distribution of the faculties in these measures.

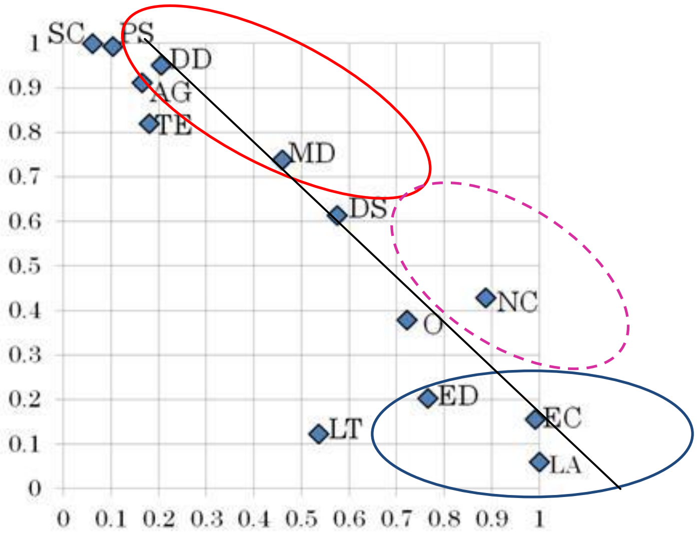

Figure 6. Distribution of Faculties with the Similarity with LA (x-axis) and with SC (y-axis).

We can see in Figure 6 that 13 faculties are divided into 3 groups; the red group at the top-left corner (i.e. low similarity with LA and high similarity with SC), the blue group at the bottom-right corner (i.e. high similarity with LA and low similarity with SC), and the purple group for those in-between the red and blue groups (i.e. medium similarities both with LA and SC).

It is very popular in Japan to classify a person who has graduated from a university into two stereotypes based on his/her major in the university; one is “Rikei” and the other is “Bunkei.” Rikei indicates that he/she is (natural) science oriented and will behave accordingly, whereas Bunkei literary means “Literature” oriented and it indicates he/she is humanity/social/art oriented. It is interesting to see that the faculties for the red group exactly match to Rikei (science oriented) and those for the blue group exactly match to Bunkei.

The expression for the faculty of DS (Design) in Japanese is actually the “Faculty of Art and Engineering” and thus the students of DS consist of those from these two types. The faculty of NC (New Century, or the 21st Century) was founded for the students who do not have or do not intend to have explicit major so that they can choose whatever to learn according to their interests and needs from the beginning. Thus a student of NC may have the mixed interest areas from Rikei and Bunkei. From Figure 6, we can see that the students of the faculties DS and NC are actually located between, or mixed up with, Rikei and Bunkei as are expected. However with having a closer look, DS is somewhat close to Rikei (red group, or natural science oriented) and NC and O (Others) are rather close to Bunkei (blue group, or social science oriented).

Based on the results in Figure 6, we choose the order of the faculties according to the value of the similarity with LA subtracted by the similarity with SC. In other words, the order is taken according to the projected point of the faculty in Figure 6 to the diagonal line segment that connects the two terminal points of (1, 0) and (0, 1).

Table 2. The Differences of Similarity with LA Subtracted by Similarity with SC together with the Order

	AG	DD	DS	EC	ED	LA	LT	MD	NC	O	PS	SC	TE
LA-SC	-0.7	-0.7	-0	0.83	0.56	0.94	0.41	-0.3	0.46	0.34	-0.9	-0.9	-0.6
Order	11	10	7	2	3	1	5	8	4	6	12	13	9

Table 2 shows the result of the faculties' difference values of the similarity with LA minus the similarity with SC and the resulting order; LA, EC, ED, NC, LT, O, DS, MD, TE, DD, AG, PS, SC in the decreasing order. We take this order here after instead of the lexicographic order we have used so far. Figure 7 shows the revised patterns of the faculties using this order. Roughly speaking, the order of the faculties, from left to right, start with those of Bunkei, followed by the intermediate, or mixed of Bunkei and Rikei, faculties, and end with Rikei faculties. Thus the similarity values for Bunkei faculties start from high values, and going down to small values, whereas for Rikei faculties the line goes up in the opposite way.

In this respect LT (Letter) is exceptional in the sense quite a lot of curves have big decreases at LT, which means LT is somewhat different from others in its profile. This specialness also appears in Figure 6: LT is the only faculty that is located far away from the line indicated in the figure. This finding is very interesting because even with the name of the faculty, i.e. "Letter," is expressed by "Bun" in Japanese, so it is the "Faculty of Bun" and thus it gives the impression that it represents the Bunkei faculties, LT is quite different from other ones.

MD (Medicine) is also located in the position which is against our intuitive image. MD students have to learn quite a big volume of knowledge from mathematics, physics, chemistry, and biology. Intuitively they have the image to be representative Rikei (natural science oriented) people. However according to the loan records they are the closest to Bunkei in the Rikei faculties. One possible interpretation to this mysterious fact is that they are more human oriented in fact because they need to know humans from a wider scope.

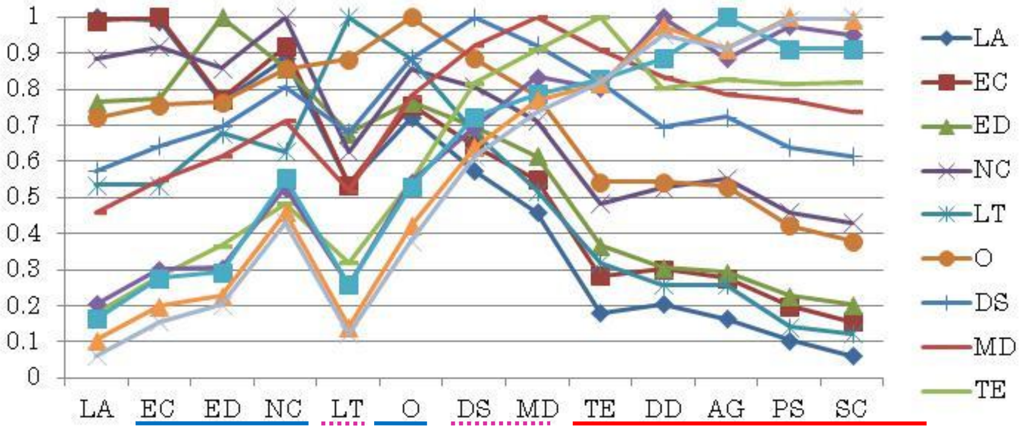

Figure 7. Similarities of 13 Faculties in Comparison with themselves

3.4 Similarity Analysis of Top 11 Students with New Faculty Order

Figure 8 shows the same data as of Figure 5 using the new faculty order from LA (left-most) to SC (right-most). It becomes clearer to see the characteristic features of students. The one who is typically Rikei, or natural science oriented, has the curve which starts with small value, i.e. lower left part of the graph, going up in the middle, and ends with big value, i.e. top right part of the graph. The one of Bunkei, or humanities oriented, has the opposite curve; starts with big value and ends with small value. Some students have different patterns. For example, P.A has the curve with moderate ups and downs; the differences of values are very small in comparison with other students. P.B has very small similarity values for most faculties except for LT.

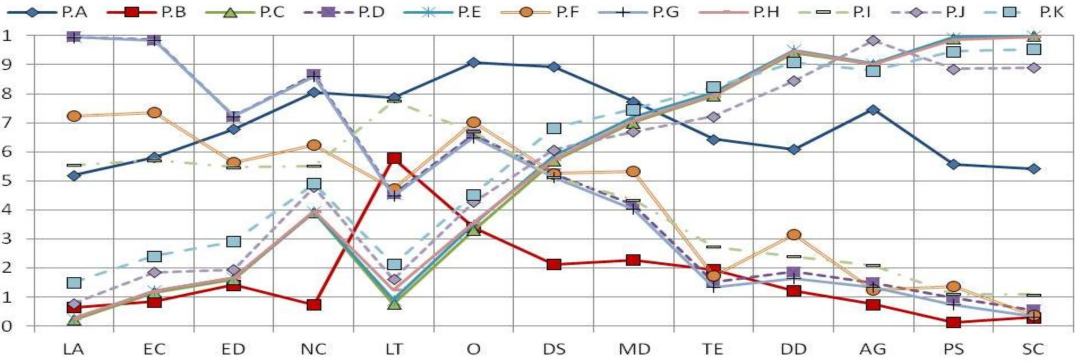

Figure 8. Similarities of 11 Patrons in Comparison with 13 Faculties (Revised)

In order to have better views of the similarities we divide the 11 patrons into 3 groups. Figure 9 shows their similarities to the faculties in 3 radar charts. We can see in Figure 9 (up-left) that the 4 students affiliated in SC (P.C, P.E, P.H, and P.K) and another one (P.J) who is affiliated in AG, have very similar patterns and thus their interest areas are not only similar in their region sizes but also in the areas themselves, even though their types vary from B3, B4, to M, and their strengths of profiles are from 185 to 148. It is interesting to see that P.J is affiliated in AG and has the highest similarity of P.J is to AG (0.98), thus we can say that P.J is a typical AG student. The similarity pattern of P.J resembles to that of SC and thus this patron's interest areas are those of natural sciences and technologies.

Among 3 students affiliated in LA in Figure 9 (up-right), P.D and P.G are very similar so that their lines are almost overlapped. Thus their ranges are similar; 0.12 for P.D and 0.10 for P.G. Their strengths are little bit different; 183 and 167. Their types are the same; D. The rest student, i.e. P.F, is quite different from these two students. P.F is also a Ph.D student (type D) and has almost the same interest strength of 168. But the preference to the areas is quite different. The range size is 0.35, so P.F prefers to read wider areas of books than the other two. Still P.F has similarity to other 2 in the sense the similarity values are very low against the faculties from PS to DD from Figure 9 (up-right). Interestingly this tendency is a kind of opposite to that of patrons of SC in Figure 9 (up-left) that the patrons of SC have relatively high values for these faculties. Thus we may say that these differences are typical ones between students of Rikei and Bunkei.

The rest 3 patrons in Figure 9 (down) have their own patterns. P.A's interest areas have relatively high similarities to most faculties, so we can guess this patron has quite a wide area of interesting topics. It is also supported by the range size value 0.95. P.B belongs to LT (Letter) and thus has quite a high similarity to LT. So we can say P.B is a typical member of LT in his or her interest areas. P.I is another one belonging to O. The similarity pattern of P.I highly resembles to that of the major LA students in the sense that P.I's similarity values are smaller, except the P.I's similarity to LT is relatively bigger than typical LA students. So we can say that P.I is virtually nearly belongs to the faculty of LA.

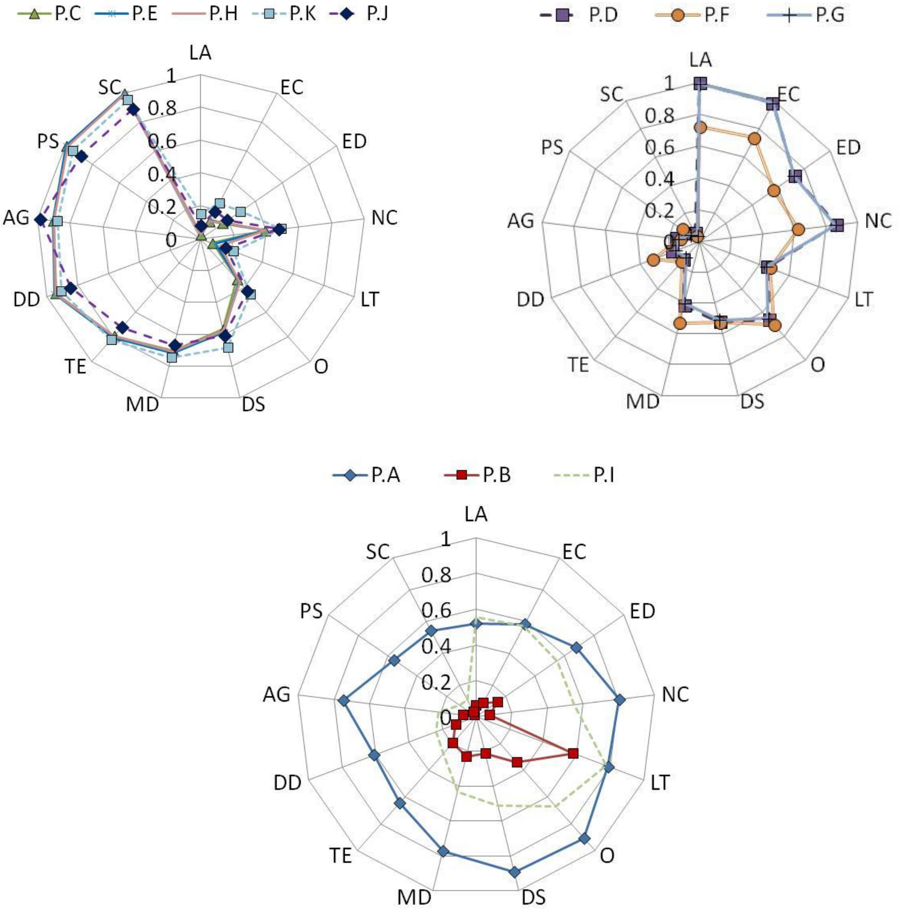

Figure 9. The Radar Chart Views of Similarities Divided into 3 Groups. (up-left) 5 Students with Typical Rikei Patterns SC, (up-right) 3 Students with Typical Bunkei Patterns, and (down) Remaining 3 Students with Mixed-up Patterns

4. ANALYSIS WITH VIRTUAL FACULTIES

Based on the observations so far we introduce the concept of virtual faculty, or affiliation, of a patron. The i-th virtual affiliation is the i-th similar faculty of the patron. We will just call the virtual faculty for the 1st virtual faculty. Then, for example, the P.A's virtual faculty is O and his/her (real) faculty is also O, and P.B also has the same real and virtual faculty LT. Among the 11 patrons from P.A to P.K, only 2 have different faculties; P.F has LA as the real faculty and EC as the virtual faculty, and P.J has O as the real faculty and LT as the virtual faculty.

As we have a closer look, we can see that P.A has the similarities against the faculties in the order of O (0.91) > DS (0.89) > NC (0.81) > LT (0.79) > … > PS (0.56) > SC (0.54) > LA (0.52). Thus P.A's 1st virtual faculty is O, followed by DS for the 2nd, and NC for the 3rd. For P.B, the faculty order becomes LT (0.58) > O (0.34) > MD (0.23) > … > SC (0.03) > PS (0.01). Even though the virtual affiliation of P.B becomes LT, the similarity value is much lower than that of P.A. So we can say that P.A is a typical member of the virtual faculty O whereas P.B is not so much typical as a member of LT.

Figure 10 shows the number of students who are affiliated with the real faculties and Figure 11 shows the number of students for the virtual faculties, from the first one to the third one. We can see that the faculty that has most students as its members is SC. The numbers are 1,168 for real and 1,818 for virtual. Considering that the profiles of SC and PS are very similar and their similarity is as high as 0.99, and still the students having PS as the virtual faculty are quite small in number most students who are concentrated in learning natural sciences (with NDC 400) are more like SC than PS. As we have a look of their profiles in Figure 2 (right), they look like the same. The difference we can see between them is that PS is less concentrated in the interest areas than that of SC in the sense PS has bigger ratios for the NDC categories of 800 (Language) and 300 (Social Sciences). Such seemingly small differences might cause the big difference of the number of students who have virtual faculties of them. So we can conclude that most students who learn natural sciences (NDC 400) are very concentrated to natural sciences so that they do not borrow books of other categories.

The second largest real faculty in the number of affiliated members is AG (848). However the corresponding number of members in the virtual AG drops largely to 312. This fact will inspire that quite a many students affiliated to AG do not have the profiles that have different patterns from the total pattern of AG. The faculties of EC, MD and O have the similar phenomena of dropping a lot of the numbers in the virtual affiliations from the real. Here again, even with the profiles of AG and DD are very close (similarity is 0.89), the number of virtual students are quite different. As we compare the profiles of AG and DD, their big differences lie on the ratios for NDC 500 (Technology and Engineering) and 600 (Industry and Commerce), where DD students have almost no interest on these topics. So we can observe that quite a lot of students who are interested in natural sciences also have interests to the subjects for the NDC categories of 500 and 600.

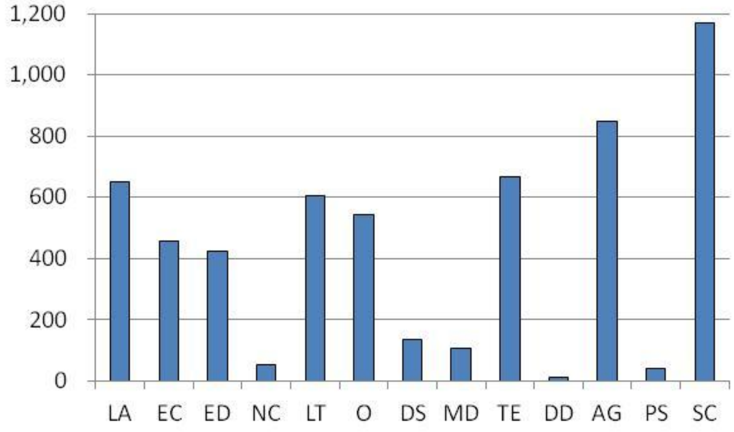

Figure 10. Comparison of the Numbers of Patrons of Real Faculties

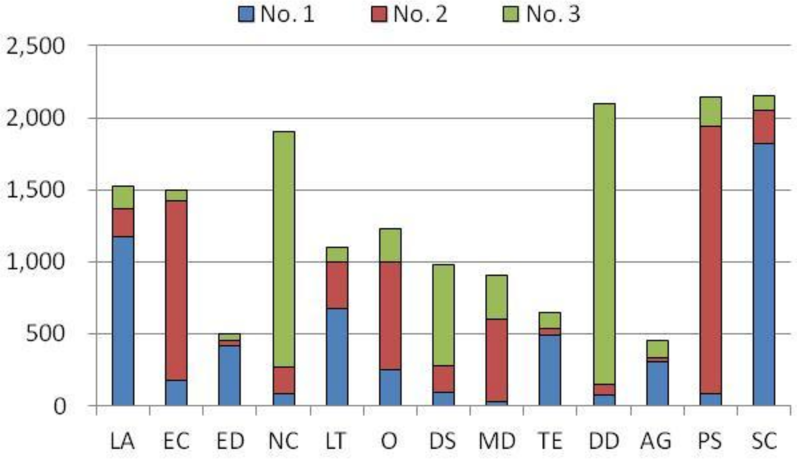

Figure 11. Comparison of the Numbers of Patrons of Virtual Faculties (from the 1st to the 3rd)

It is interesting to see that some faculties have quite many members as the 2nd and 3rd virtual affiliations. For example, PS, EC, and MD have quite a lot of members as the 2nd virtual faculties. Similarly, DD, NC, and DS have several times as many new patrons as their 3rd virtual faculties. We have no explanation yet on this phenomenon, and thus we need further investigation on this topic.

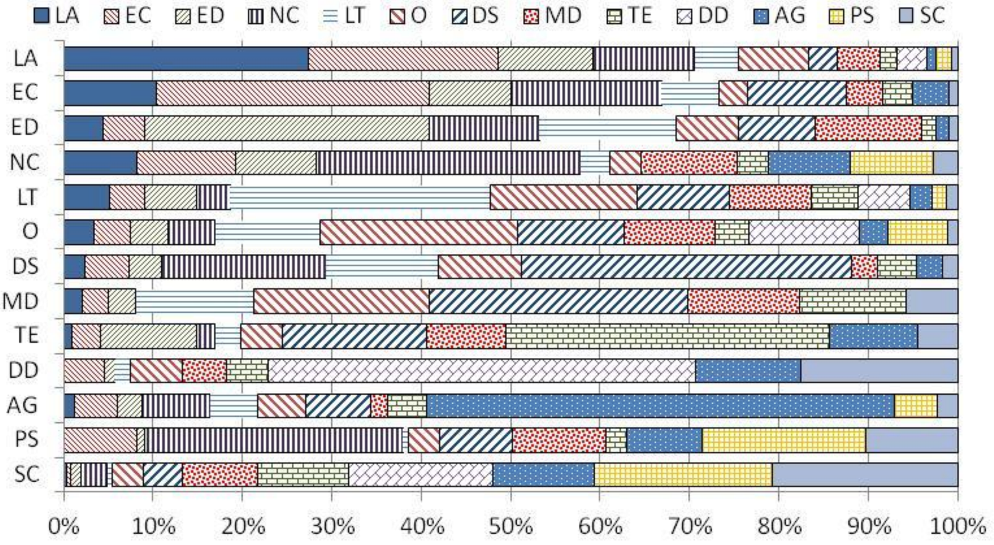

Figure 12. Ratios in Percentage of the Numbers of Virtual Faculties' Member for Each Real Faculty

Figure 12 shows the ratios of the numbers of the patrons according to the virtual faculties for each real faculty. From this figure we can see that most (81%) members of SC have SC as their virtual faculty. PS has the similar tendency that 78% of the members have SC as their virtual affiliation. DD is kind of similar to these because 64% of the members have SC as their virtual affiliation. On the other hand only 34% of the members of TE have TE as their virtual affiliation, which is smaller than the ratio for SC as virtual faculty, which is 40%. In this respect MD is the most extreame case. Only 1% of the members have MD as their virtual affiliation. Among them 33% belong to SC and 14% to LT in terms of their virtual affiliations.

5. CONCLUDING REMARKS

Our eventual goal is not only to analyze the data from library and lectures and discover knowledge that is useful in giving better educational environment to students, but also to develop the more sophisticated tools for data analysis in this respect. As a primitive step toward such a goal we have defined the interest area profile. We started with discussing the importance of such approach in improving education. Then we defined the profile of a patron from interest area, and we define the concepts of strength and the range of the profile, and thus of the patron. We extended these definitions to the profile of a group of patrons.

After these preliminary investigations, we took the top 11 students as samples and started analyzing their interest areas through comparing the similarities with those of faculties and obtained some interesting results as an extension to our previous study in [9]. In Section 3, we introduced a similarity measure and we have found some number of interesting results. In Section 4, we introduced the concept of the virtual affiliation/faculty of a patron and demonstrate its usefulness in understanding the patron's interest areas in comparison with those of the faculties as a case study. The relationships between real and virtual faculties of students may be able to use to characterize a university and its students, which can be a new approach to measuring a university.

Because of the usefulness of loan records, they are used in evaluating of library collections [3]. They are usually analyzed with various kinds of statistical methods and to efficiently recognize the representative image of the total data. The system WorldCat Collection Analysis system [10], for example, provides an easy-to-use analysis environment to librarians, based on such standard statistical methods. A research on loan record analysis for evaluating the usage of e-books is reported in [3]. In addition to these research based on the statistical methods, investigation of the association rules in classification category of books using a data mining method is reported in [1]. In our previous works [4-6] we defined the concept of expertise levels of books and patrons and investigated the expertise levels of faculties. Our methodology to library data analysis is applicable also to lecture data, which are another source of data for analysis [7]. Goda et al. developed a method for analyzing lecture data in a similar approach to ours [2].

Our future plans include (1) to investigate more about the characteristic features of a patron and a group of patrons, (2) to develop different concepts for measuring, indexing, characterizing some behaviors of a patron and of a group of patrons, and (3) to extend our research area in order to cover wider area by introducing different types of data including lecture data, other types of educational data, etc.

Even though our research direction is quite new and we have little studies that have similar methodologies so that the research level in this field is still very primitive. Furthermore it is difficult to obtain the library data due to privacy issues and thus we have to put more effort to demonstrate the usefulness of our approach in library marketing. Even with such difficulties, we are convinced from our experiences so far that out approach has high potential and thus it will create practical results in near future.

ACKNOWLEDGEMENT

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] Cunningham, S.J. and Frank, E., 1999. Market basket analysis of library circulation data. *Proc. 6th International Conference on Neural Information Processing*. Perth, Australia, pp. 825-830.

[2] Goda, K. and Mine, T., 2011. Analysis of Students' Learning Activities through Quantifying Time-Series Comments, *Proc. 15th Annual KES Conference (KES'2011), Part II, Lecture Note in Artificial Intelligence (LNAI 6882)*. pp. 154-164.

[3] Littman, J. and Connaway, L.S., 2004. A Circulation Analysis of Print Books and e-Books in an Academic Research Library. *Library Resources & Technical Services*, 48(4). pp. 256-262.

[4] Minami, T., 2012. Book Profiling from Circulation Records for Library Marketing -- Beginning from Manual Analysis toward Systematization --. *International Conference on Applied and Theoretical Information Systems Research (ATISR 2012)*. Taipei, Taiwan, 15pp.

[5] Minami, T., 2012. 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*. Fukuoka, Japan, pp. 357-362.

[6] Minami, T. and Baba, K., 2012. Investigation of Interest Range and Earnestness of Library Patrons from Circulation Records, *International Conference on e-Services and Knowledge Management (ESKM 2012), as a part of the 1st IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), IEEE CPS, DOI 10.1109/IIAI-AAI2012.15*. Fukuoka, Japan, pp. 25-29.

[7] Minami, T. and Ohura Y., 2012. Towards Development of Lecture Data Analysis Method and its Application to Improvement of Teaching, *Proc. 2nd International Conference on Applied and Theoretical Information Systems Research (2nd ATISR 2012)*. Taipei, Taiwan, 14pp.

[8] Minami, T., 2013. Profiling of Patrons’ Interest Areas from Library’s Circulation Records – An Approach to Knowledge Management for University Students --. *The Fifth International Conference on Information, Process, and Knowledge Management (eKNOW 2013)*. Nice, France, pp.6.

[9] Minami, T., 2013. Interest Area Analysis of Person and Group Using Library’s Circulation Records. *IADIS International Conference on Information Systems 2013 (IS 2013).* Lisbon, Protugal. 8pp.

[10] Online Computer Library Center, Inc. (OCLC). WorldCat Collection Analysis. http://www.oclc.org/collectionanalysis/