Profiling of Patrons' Interest Areas from Library's Circulation Records – An Approach to Knowledge Management for University Students –

Toshiro Minami
Kyushu Institute of Information Sciences,
Dazaifu, Fukuoka, Japan, and
Kyushu University Library,
Fukuoka, Japan
Email: minami@kiis.ac.jp, minami@lib.kyushu-u.ac.jp

***Abstract*—The concept of knowledge management is important not only in industries but also in educational organizations like universities. Considering the importance of this concept, it is not surprising that many universities have introduced the database system for saving the profiles and history information of students and utilize them in order to improve their educational abilities. In order to make the information more effective in education, it is preferable to collect not only the information in raw but also the knowledge that is found by the data with data analysis and data mining. In this paper, as a new approach to knowledge mining in education as a part of educational knowledge management, we deal with the circulation records of a university library as the target data. The library's circulation records show the relationship between the patrons and the books, which are usable to know the patrons about their fields of interest, knowledge levels, and other information. In this paper, we put special emphasis on the investigation of the profiling of students as a knowledge management. As a part of this, we deal with the interest area of a student and explore the measuring methods for the profiling of the student patrons.**

***Keywords*-knowledge management; knowledge discovery; library marketing; data analysis; data mining;**

I. INTRODUCTION

The most important mission of a university as an educational organization is to provide its students with good learning environment. The concept of knowledge management is important not only in industries but also in such an educational organization so that it can manage the data concerning its students' learning ability, willingness to study, and other aspects that relate to learning and studying.

Considering the importance of this issue, it is getting to be popular and many universities have introduced the database systems for saving the profile and history information of students and utilize them in order to improve their educational abilities. Such a database is sometimes called a learning portfolio, student portfolio, or digital portfolio, etc.

In order to make the information more effective in education, it is preferable to collect not only the information recorded by teachers and other staff as an original data but also the knowledge that is found by the data with data analysis and data mining.

In this paper, as a new approach to knowledge mining in education, we deal with library data analysis, especially the circulation records. The library's circulation records have an advantage because every university has a library network and every library must have the circulation records as necessary data in their services. In this paper, we take the circulation data of a university library as the target for analysis and show how to extract useful information out of them. The library's circulation records deal with the relationship between the patrons and the books, which can be used to know the patrons about their fields of interest, knowledge levels [6], and other information.

So quite a lot of researches have been conducted so far. For example circulation records are used for evaluation of collections of library in [2]. They are usually analyzed with various kinds of statistical methods, which are very useful to efficiently recognize the representative image of the total data. The system WorldCat Collection Analysis [11], for example, provides an easy-to-use and easy-to-recognize analysis environment to librarians, based on the standard statistical methods. A research on circulation record analysis for evaluating the usage of e-books is reported in [2]. Yamada analyzed the circulation records of a university library with considering the material age of the circulated books [12]. 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].

Our approach to circulation record analysis is different from such standard methods. We take the analysis methods combining two ways; one is the statistical one for surveying the general tendencies of the patrons, and another one is the new way trying to find the more realistic patron's behavioral model and to understand the patrons' behavior in reading, studying, and using of libraries more precisely, including their underlying needs, preferences etc. In this paper we show some example data analysis experiences as a case study using the circulation records of the Central Library of Kyushu University, Japan (KUL) for the academic year 2007. The results shown in this paper are extensions to the

 ISBN: 978-1-61208-254-7

results reported in [3], [4], [5], and [8]. Our aim in this paper is to propose new methods for getting more precise patron profiling as a whole and a patron's preference, knowledge level, eagerness to learning, etc. that will be helpful for personalized services in learning assistance.

In the paper [6], we proposed the concept of p-rank for measuring expertise level of a book and of a patron. In this paper we propose two new concepts for measuring interest range size and earnestness in learning. We compare faculties in their features by applying these measures.

The rest of this paper is organized as follows: In Section II we review our previous research results and show some case studies that inspire the research presented in this paper. In Section III, we propose a concept of the profile of a patron that indicate the patron's interest. Then we define two concepts for measuring interest range size and strength of earnestness for studying from the patron's profile. We then investigate how to capture the patrons behavior through these measures. We also define the similar concepts for a group of patrons and apply them to the comparative study of faculties. Finally in Section IV, we summarize what we have done in this paper and prospect possible future works.

II. Profiling of Patron with Data Analysis from Library Data

This section describes some of our case studies on data analysis of library data. The study in the next section is an extension to these analysis experiences.

A. Target Data for Profiling with Data Analysis

In this paper, we use the circulation records obtained in the Central Library of Kyushu University, Japan, in the academic year 2007; i.e. from April 2007 to March 2008. The whole data contain 67,304 circulation records. A record item consists of the book ID, book's classification number, call number, borrower's patron ID (renumbered one so that the record does not link to the real patron ID), borrower's affiliation, borrower's type (undergraduate student, masters student, Ph.D student, professor, staff, others), and the timestamps for borrowing and returning, etc.

The number of patrons, who borrowed at least one book during this year period, is 6,118 in all and the average number of borrowed books per patron is about 11.

A circulation record has 10 patron types: undergraduate student (Bachelors-1 to 6, or B1 to B6), masters student (M), Ph.D students (D), academic staff (Professors, P), and others (O). About 45% of books are borrowed by undergraduate students and 24% by masters and 15% by Ph.D students. Thus about 80% of books are found to be borrowed by students; which supports based-on the objective data that the frequently-told saying that most important patrons of university libraries are students.

B. Preprocessing of Circulation Records

As a preprocessing, we eliminate the records that have inappropriate values and no data for the inevitable properties (items) that are necessary to deal with in the analysis in this paper. For example 244 records have NDC (Nippon Decimal Classification) numbers that are greater than 1000 and 7,260 records have the non-numeric values for this item and thus have eliminated from the original records. After elimination, 53,182 records are left as those for analysis.

C. Case Study: Expertise Level as a Profile for Library Patrons and Library Books

The concept of the expertise level of a patron is useful in various purposes in such cases as to recommend books to read, to form a study group, to estimate the period of times to need for the patron to study some specific subject, etc. We defined an expertise level measure of a book and a patron, which we call p-rank in both cases [6].

We defined the expertise level of a book as the average value of its borrowers' initial expertise levels; where the initial expertise levels of B1 to B6 are set to 1 to 6, respectively, 8 for M, 9 for D, and 10 for P. We do not count the patrons of the type (O). Then we define the expertise level of a patron as the average expertise levels of the books the patron borrows in the circulation records. See the paper [6] for more detail about p-rank, and c-rank, which is another definition for expertise level of a book.

D. Definition of a-value as Another Measure for Expertise Level

As another idea for defining expertise level of books with assuming that if a book is borrowed by a limited number of patrons then its expertise level is high. In other words, if a book is borrowed by a wide range of patrons, its expertise level is low.

Based on this assumption we define the a-value (affiliation based expertise level) of a book [5]. Firstly we have to choose the faculties as the representatives of expertise fields. Affiliations of Kyushu University consist of not only the faculties for undergraduate students but also of some number of research centers, library, communications center, and others. We will take 12 faculties together with the graduate schools for graduate students relating fields and research organizations for professors; precisely, SC for (Faculty of) Sciences, AG for Agriculture, TE for Engineering, MD for Medicine, DD for Dental, PS for Pharmaceutical, LA for Law, LT for Letter, EC for Economy, ED for Education, DS for Design, and 21 for the special faculty of Kyushu University called 21st century program, which was founded for the students who are willing to study from a wide variety of learning fields.

So there are 12 groups based on the faculties. Let m be 12 as the number of categories and let F_i $(i = 1, 2, \ldots, m)$ be the i-th faculty. The a-value of a book is calculated as

 ISBN: 978-1-61208-254-7

follows. Let CR be the set of circulation records; $CR = \{r =< BookID, NDC\ Number, Borrower, Borrowed\ Day\ and\ Time, Returned\ Day\ and\ Time, \ldots >\}$. We use $BI(r)$ for the book ID, $Cls(r)$ for the NDC number, $B(r)$ for the borrower, $Bd(r)$ for the borrowed day and time, and $Rd(r)$ for the returned day and time, of r. For a given book b, let us define the number s_i $(i = 1, 2, \ldots, m)$ of the book for the faculty F_i by $s_i = \#\{r \in CR | BI(r) = b, B(r) \in F_i\}$, where $\#$ is the number of elements. We put 0 for a-value if all the s_i's are zero; i.e. that the book is borrowed by the patrons who do not belong to these faculty related affiliations. Let us set $s = \Sigma_{i=1}^{m} s_i$, i.e. total number of circulations borrowed by the patrons belonging to either one of the nominated faculties. Then we define the a-value of the book b as $10 \times \Sigma_{i=1}^{m}(s_i/s)^2$, where multiplication value of 10 is used in order to make the maximum value to 10 so that it becomes easier to compare it with other values.

III. Interest Area Analysis from Circulation Records

A. Profiling Interest Area of Patrons

The eventual goal of the study in this paper is to provide library patrons with good learning environment. Mostly the services provided by libraries are intending to be universal; to every patron in a uniform way and thus in a uniform level. However toward the future, personalized services are expected to be more and more important for libraries. With personalized services, patrons are able to get better assistance that matches more to the patrons' needs and will have better effects in learning. In order to provide with unique services we would like to investigate in developing methods of analyzing the library data and to obtain knowledge about the profiles of patrons.

In this paper, we deal with profiling the patrons' interest areas by analyzing circulation records. The concept of interest areas of a patron may be considered good for characterizing the patron's attitude to learning. We would be able to extend the profile on interest areas to other properties of patron that relate more on knowledge level, learning abilities, learning styles, etc.

We use the classification field of book using the NDC number of the book. NDC is a decimal classification system like DDC (Dewey Decimal Classification) localized to Japan. The top level categories consist of the following 10 topics; 000 for General Works, 100 for Philosophy and Religion, 200 for History and Geography, 300 for Social Sciences, 400 for Natural Sciences, 500 Technology (Engineering), 600 for Industry and Commerce, 700 for Arts, 800 for Language, and 900 for Literature. Note that NDC classification items are different from those of DDC.

For a patron p, we define the profile $Prof(p)$ of p as the vector of frequencies of the books borrowed by the patron p according with the books' 10 classification numbers from 000 to 900 in NDC.

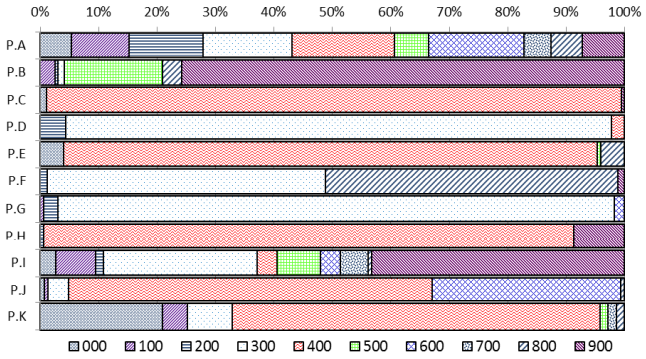

Figure 1. Profiles of the Top 11 Patrons in the Numbers of Borrowed Books, or Items

For a patron p, we define $Prof(p) = \{< bt(c) >_{c \in NDC} \mid bt(c) = \#\{r \in CR \mid B(r) = p, \text{ and } Cls(r) = c\}\}$.

We apply this definition of $Prof(p)$ to the 53,182 circulation records that are described in Section II-B. Figure 1 shows the profile patterns in 100% stacked column chart of top 11 patrons in terms of the number of borrowed books. We will call them from Patron A (P.A) to Patron K (P.K) in the order of the numbers of borrowed books; from 388 by P.A, 268 by P.B to 143 by P.J and P.K.

It is easy to see 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 number from 0 to 9. On the other hand, P.C borrows mostly with the classification number 400 (Natural Science). Such difference about the range 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 characteristic features of a patron, which will be discussed in more detail in the next section.

B. Interest Range and Interest Strength Measures

In order to analyze deeper about the interest profiles of patrons, we define 2 new measures; strength and range of interest of a patron.

The interest strength is a concept intending to measure the willingness to learn and to know, or earnestness to obtain new knowledge. We use the number of books, or items, that the patron borrows as the measure for interest strength; $Str(p) = \#\{r | r \in CR, B(r) = p\}$ for a patron p.

Interest range is also a very important measure to describe about the willingness of learning of a patron. As we see Figure 1, we can easily recognize that P.A is interested in quite a wide areas of topics, whereas P.C is mostly interested in one subject only. In order to compare such difference of the patterns of interest we propose a new measure for the amount of the width of interest of patron. We use the concept of entropy, or the amount of information, for the interest range of a patron. Let p be a patron. We define the

 ISBN: 978-1-61208-254-7

Table I
COMPARISON OF PATRONS IN THEIR PROPERTIES

Patron	Range	Strength	Affiliation	Type
P.A	0.95	388	O	O
P.B	0.34	268	LT	D
P.C	0.04	185	SC	B4
P.D	0.12	183	LA	D
P.E	0.16	173	SC	B3
P.F	0.35	168	LA	D
P.G	0.10	167	LA	D
P.H	0.15	150	SC	B4
P.I	0.72	148	O	M
P.J	0.38	143	AG	B3
P.K	0.49	143	SC	M

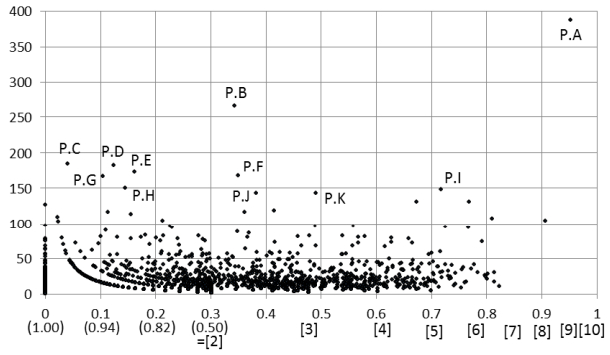

Figure 2. Correlation between the Range (x-axis) and the Strength (y-axis) of All Patrons

(interest) range of p as follows:
$Range(p) = \sum_{Prof(p)_c \in NDC}(\frac{Prof(p)_c}{Str})log(\frac{Prof(p)_c}{Str})$
where $Str = Str(p)$ and $Prof(p)_c$ is the number of books borrowed by the patron p having the NDC number c among $NDC = \{000, 100, 200, \ldots, 900\}$. We take 10 for the base of the logarithm so that the maximum value of the range becomes 1 because the number of the categories, i.e. number of the NDC values, is 10.

Table I shows the range, strength, affiliation, and type of the 11 patrons from P.A to P.K. As has been predicted the range of P.A (0.952) is quite high; the highest among 11 patrons. On the other hand P.C has the minimum range value (0.04), who's affiliation is SC and the year 4 undergraduate student (B4).

To have a closer look at the table, there are 4 students with affiliation of SC (Sciences) and 2 of them are B4 (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 just like P.C. It is interesting to see that the remaining masters 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); 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. 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 2 shows the correlation between the range size (x-axis) and the strength (y-axis) for all patrons. 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 from each category. The location of the numbers parenthesized with [n] indicate that it is the range value, i.e. entropy, for the case that n categories have equal numbers of books, or $log_{10}n$, which is the maximum value for having values in n categories. The location of the numbers parenthesized with (n) indicate that it is the range value for 2 categories in which one category has the possibility of n and the other has the possibility of $1 - n$. In this case, the maximum range value is $log_{10}2 = 0.30$ when n=0.5, i.e. half and half.

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 graph. Patron A (P.A) 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 is over 100 books more than the second one, i.e. P.B, who borrows more than 250 books.

The patrons C, D, E, G, and H are located in the left-most part of the graph having the value 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 B, F, J, and K are located in the range with the range value from 0.3 to 0.5, which means, roughly speaking, they mainly borrow books with 2 or 3 categories.

C. Interest Profile for a Group of Patrons

The definition of profile of a patron is naturally extendable to a group of patrons. Let P be a group of patrons, then the profile of the group P is defined as follows:
$Prof(P) = \{< bt(c) >_{c \in NDC} \mid bt(c) = \#\{r \in CR \mid B(r) \in P, \text{and } Cls(r) = c\}\}$. In other words, the c-th component of $Prof(P)$ is the sum of the c-th component of the members of the group P; $Prof(P)_c = \sum_{p \in P} Prof(p)_c$.

Figure 3 shows the profiles of the affiliations of patrons. The names in the figure come after the faculty names; the graduate students and academic staff's affiliation names are assigned to the faculty names that are mostly closed to the patrons' affiliations. Patrons who have no appropriate relationship to a faculty is assigned to other (O).

 ISBN: 978-1-61208-254-7

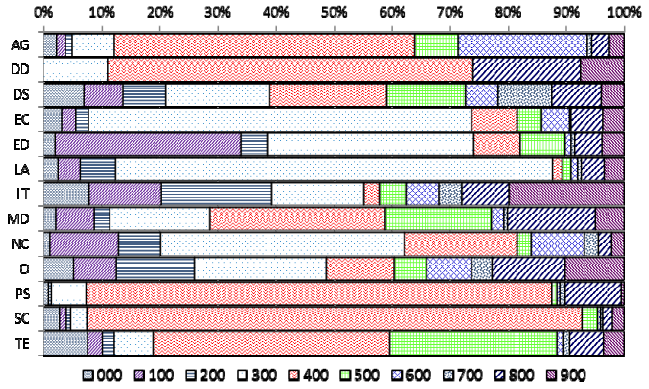

Figure 3. Profiles of Patrons' Affiliations (or "Faculties")

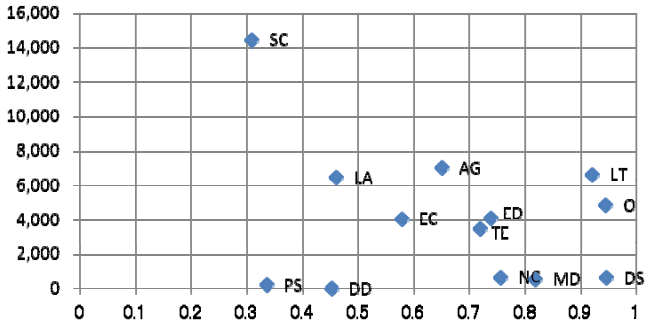

Figure 4. Correlation between Region Size (x-axis) and Strength (y-axis) of Faculties

Figure 4 shows the correlation between region size and strength of faculties. SC (Sciences) is far away from other faculties in both axes. It has the lowest value in region size and the highest in strength, which mean 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 because that their places locate very close to the library and thus it is quite easy for them to visit the library and borrow many books.

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 strengths, or the numbers of borrowed books, of PS and DD is that their faculties locate in a different campus from where the library locates. Thus the patrons in PS and DD visit the library in order to get the books they could not find in the libraries in 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 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.

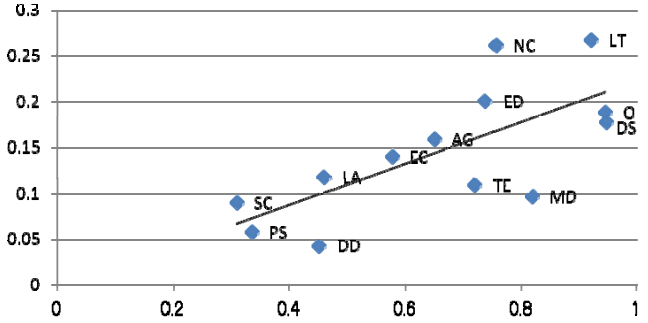

Figure 5. Comparison between the Range Size (x-axis) and the Average Range Size of the Members (y-axis) of Faculties

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 that of 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 a surprising fact that its range size is larger than any other faculties including O (Other, or unclassified).

Another surprise is that LT (Letter) has high range size. LT patrons borrow books not only of literature (NDC 900), but also of those in other areas nearly as many as of literature.

We are able to define the concept of interest range of a group, or faculty in the current situation, in another point of view; the average range size of the members of the group. Formally, $AverageRange(P) = average\{Range(p) \mid p \in P\}$ for a group P of patrons. Figure 5 shows the difference of the range size and the average range size of faculties. The line in the graph is the linear approximation line.

From the definition, we can see that even if each member have low range size, that as a group is much wider if each member's interest area is different. In other words, if all the members of a group have exactly the same profile and thus have the same range size, the average range size is the same value as of members. So we can say roughly that the difference between the range size and the average range size indicates the varieties of the interest profiles of the members.

From this view, or interpretation, quite many faculties are close to the approximation line and thus they have average interest varieties of the members. The faculties LT and NC have larger average range size than the one on the line. So we can say that these faculties have a wider variety of members in terms of interest ranges.

On the other hand DD, MD, and TE locate in the lower area of the line. So we can say that the members in these

 ISBN: 978-1-61208-254-7

faculties have a narrower variety of interest profiles than other faculties; i.e. the members have somewhat similar interest ranges.

IV. Concluding Remarks

We proposed a new concept of interest area profile of a patron using a set of circulation records of a university library. Considering that one of the most important missions of a library, especially of a university library, is to help its patrons with learning more effectively and more efficiently in the comfortable and enjoyable environment. In order to achieve this goal, capturing the profiles of patrons in terms of their learning styles, their learning histories, their knowledge levels, their interests, their preferences, and so on, is important.

In order to compare the profiles of patrons in more practical ways, we additionally proposed new concepts of strength and range (size) of the profile. The strength intends to represent the eagerness or diligence to learning of the patron and we take the number of the borrowed books, or items, of the patron in this paper. The range size intends to represent the amount of eagerness or earnestness of the patron in terms of the width of the topic areas. We took the information entropy for defining this concept in this paper. We see the patrons' characters not only with profiles but also with these two values.

The concept of profile was extended to a group of patrons, especially to faculties. The concepts of strength and range were also extended to groups. We compared and characterized the faculties by these concepts and analyzed the characteristic features of faculties.

Our approach to library data analysis is quite new and there are no other such studies to our knowledge. Even though our current analysis methods are still in a primitive level, we are convinced from our experience that our methods have high potential as a tool for library marketing, and thus it will become an essential tool in the future.

The research directions of this paper include the topics:

- Investigation of more appropriate definitions of the concepts of the amount of eagerness to learning and the interest range size
- Exploration of defining other concepts such as style of learning, learning pace, preference in learning, etc. of a learner
- Utilization of circulation records and other data that are obtainable by libraries
- Usage of other data from different sources; for example usage of lecture data
- Systematizing the analysis methods and developing a learning support system and/or knowledge management system

This research was partly supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 24500318, 2012.

References

[1] S. J. Cunningham and E. Frank, Market basket analysis of library circulation data, Proceedings of 6th International Conference on Neural Information Processing, 825-830. IEEE Computer Society, Perth, WA, Australia, 1999.

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

[3] T. Minami and E. Kim, Data Analysis Methods for Library Marketing, The 2009 International Conference on Database Theory and Application (DTA 2009), LNCS, 5899, 26-33, Springer, Heidelberg, 2009.

[4] T. Minami, Challenge toward Patron Understanding - A Search for Patron's Profile through Circulation Data of Library –, Kyushu University Library: Annual Report 2010/2011, 9-18, 2011. (in Japanese)

[5] T. Minami, 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), pp.15, 2012.

[6] T. Minami, 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.

[7] T. Minami and Y. Ohura, Toward Learning Support for Decision Making – Utilization of Library and Lecture Data –, 4th International Conference on Intelligent Decision Technologies (KES-IDT'2012), Springer Smart Innovation, Systems and Technologies 16, pp.137-147, 2012.

[8] T. Minami and K. Baba, 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, pp.25-29, 2012.

[9] T. Minami and Y. Ohura, An Attempt on Effort-Achievement Analysis of Lecture Data for Effective Teaching, The 2012 International Conference on Database Theory and Application (DTA 2012), T.-h. Kim et al. (Eds.): EL/DTA/UNESST 2012, CCIS 352, pp. 50–57. Springer, 2012.

[10] T. Minami and Y. Ohura, Towards Development of Lecture Data Analysis Method and its Application to Improvement of Teaching, 2nd International Conference on Applied and Theoretical Information Systems Research (2ndATISR 2012), 2012. (to appear)

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

[12] S. Yamada, Analysis of Library Book Circulation Data: Turnover of Open-shelf Books, Journal of College and University Libraries 69, 27-33, 2003. (in Japanese)

 ISBN: 978-1-61208-254-7