IADIS International Journal on WWW/Internet
Vol. 12, No. 1, pp. 65-80
ISSN: 1645-7641

KNOWLEDGE FIELD RE-CATEGORIZATION TO TUNE THE DECIMAL CLASSIFICATION SYSTEM OF LIBRARY -- AN APPROACH FROM LIBRARY DATA ANALYSIS --

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

Kensuke Baba. *Kyushu University Library. 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581 Japan.*

ABSTRACT

Along with the development of ICT (Information and Communications Technology) and popularization of mobile devices, the information environment of our society has radically changed. As a result, the library's role model also needs to be changed. The most important mission of a library is to collect materials, mainly printed books, and prepare to provide library patrons with them upon request. In order to carry out this mission, classification codes are attached to library materials, and will be shelved according to the code so that searching of specific material becomes easy. In most libraries, Decimal Classification (DC) system is chosen as their classification code. For example in Japan, libraries use NDC (Nippon, or Japan, Decimal Classification) system as their de facto standard classification system. One of the biggest problems of DC is that it is quite hard to adjust the knowledge field code according to the change of important concepts and terms as time goes. For example, computer/information science is a relatively new field for library and it was hard to find an appropriate code for it, and eventually it is assigned to the category code 007 in the third level of categorization of NDC. In this paper, we propose an index for measuring similarity between NDC categories, which reflects how much amount of the books of comparing categories are borrowed by library patrons in common. In this way NDC categorization system can be tuned according to the current status of patrons' interest tendencies. Also, we demonstrate its usefulness by comparing the interest area profiles between the one using original NDC and the new, or virtual, NDC. By using the virtual NDC, we can recognize the "real" interest area profiles of a patron or a group of patrons from loan record analysis. The studies in this direction will boost up advancing more useful library services toward the future.

KEYWORDS

Library Marketing, Library Data Mining/Analysis, Knowledge Management, Decimal Classification System.

1. INTRODUCTION

Along with the development of ICT (Information and Communications Technology) and popularization of mobile devices, the information environment of our society has radically changed so that we enjoy the benefits of information society as a result. Due to such changes, the library's social role so far does not meet anymore now with the people's needs and thus the library's service/role model needs to be changed. Therefore it is a must for libraries to develop new information services that fit to the needs in this information society as well as to enrich the traditional services they have been providing so far.

The term "digital library" has been considered as it provides with digital contents such as those of the digitized files of rare and precious books of the library. In this paper, we would like to put a strong attention to the concept of "e-library" rather than "digital library" for the services that are provided on the Internet. Among the various types of e-library services, the most important ones the libraries are supposed to empower are those of personalized services such as to be provided in the framework often called as "MyLibrary" service [1, 3].

In order to provide with good personalized services, it is quite important to capture the behaviours and the attitudes of library patrons as the profiles such as interest knowledge/subject fields, knowledge levels, preferences, etc. By collecting and analyzing such profile data, the library is able to provide with more helpful personalized services to its patrons. The resulting knowledge will be helpful also for the library itself in other management-related activities such as book collection, management decision making, staff arrangement, etc.

Interest area of a patron is supposed to be a very useful candidate as a profile because it should reflect the patron's behaviour of borrowing books, which is a core service for a library. Especially we can define the concept of patron's interest area from the loan records of a library, which are obtainable for every library and thus the method is potentially easy to implement for every library (See also the studies [2, 5, 6-8, 10-12] of loan/circulation data analysis). We will define the profile of a patron about his or her interest area as the ratios of the borrowed books in terms of the top (or main) category of decimal classification (DC) system, which is popularly used in libraries. We use NDC (Nippon Decimal Classification) system in this paper as the DC system.

One of the fundamental problems in using the loan records for patron's interest area profiling is its precision. For example, the category (or section) 007 of NDC indicates the subject field of "Information/ Computer Science" as a subcategory in section level of the main category 000 for "General Works." Other subjects relating to the Information/Computer Science are supposed to be 400 (Natural Sciences) and 600 (Industry and Engineering). One reason of such a phenomenon happens is that the newly created knowledge field such as Information/Computer Science has no appropriate level of category numbers available and thus such a field had to be squeezed into the secondly appropriate category code.

The aims of the study in this paper are:

(1) to investigate what category numbers in low level are closely related to which main categories among the 1000 section categories from 000 to 999,

(2) to propose a re-organization, or re-categorization, method of the section level NDC numbers to one of the 10 main categories according to the closeness of them,

(3) to compare the interest area of a patron or a group of patrons using the original and the modified (re-organized/categorized) NDC category grouping, etc.

as an extension of the study of the article [12].

The rest of this paper is organized as follows: In Section 2, we describe our idea of using the decimal classification category as the representation for knowledge/subject fields. In Section 3, we take a collection of loan records of a library and investigate the relationships between section level NDC categories and the main level categories using loan records. In Section 4, we compare the interest area profiles of the library patrons and groups of patrons in two ways; one of using the original NDC categories and another of using the reorganized, or virtual, categories. Finally in Section 5, we summarize our studies in this paper.

2. DECIMAL CLASSIFICATION CATEGORIES TO REPRESENT THE SUBJECT/KNOWLEDGE FIELDS FOR LIBRARIES

We take the decimal classification (DC) categories as the representations of knowledge field in this paper. DC system, e.g. DDC (Dewey Decimal Classification) system, is very popular among libraries world-wide as the classification system for their collected books. Considering that the libraries, except some special-purpose ones, collect books and materials from a very wide area of subjects, a DC system is supposed to cover all the subject/knowledge fields we can think of. Thus it is reasonable to use a DC system as a representation of knowledge field.

Among the DC systems, we choose NDC (Nippon, i.e. Japan, Decimal Classification) [13] 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 in this paper. NDC, as well as a DC in general, is a hierarchical system where the top/main level consists of 10 categories from 000 to 900, and each category is normally divided into another 10 subcategories (divisions), followed by 10 sections, and such sub-categorization process continues several times; something like 413.52 for a book on general textbook for complex analysis in mathematics. Table 1 shows the 10 main categories together with some example categorizations of the 2nd level for 300 (Social Sciences) and the 3rd, or section, level for 410 (Mathematics) that is under 400 (Natural Sciences).

Table 1. Sample NDC Categories; (left) Main/Top level, (middle) Sub/2nd level, (right) Section/3rd level.

Main Categories	Example Subcategories	Example Section Categories
000 General Works	300 Social Sciences	410 Mathematics
100 Philosophy and Religion	310 Political Science	411 Algebra
200 History and Geography	320 Law	412 Number Theory
300 Social Sciences	330 Economics	413 Analysis
400 Natural Sciences	340 Public Finance	414 Geometry
500 Technology/Engineering	350 Statistics	415 Topology
600 Industry and Commerce	360 Society	416 --
700 Arts	370 Education	417 Probability/Statistics
800 Languages	380 Customs/Folklore/Ethnology	418 Numerical Calculations
900 Literature	390 National Defense/Military Science	419 Japanese/Chinese Mathematics

3. REORGANIZATION OF NDC USING LIBRARY'S LOAN RECORDS

In this section, we deal with a collection of loan records of a library and provide with the fundamental measuring indexes for capturing the closeness of a patron and an NDC category (nb), between two NDC categories (np), and an NDC category and a top-level, or main, NDC category (npM).

3.1 Overview of the Loan Records

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). A record consists of the book ID, book's NDC category number, call number, borrower's patron ID, or PID (renumbered ID by considering the privacy issue of the patron), affiliation, patron type, and the timestamps for borrowing and returning dates and times, and other items we do not use in this paper. After the preprocessing by eliminating the records that lack necessary information for this study, 53,329 records are left. The number of the patrons/users who appear in the records is 6,173.

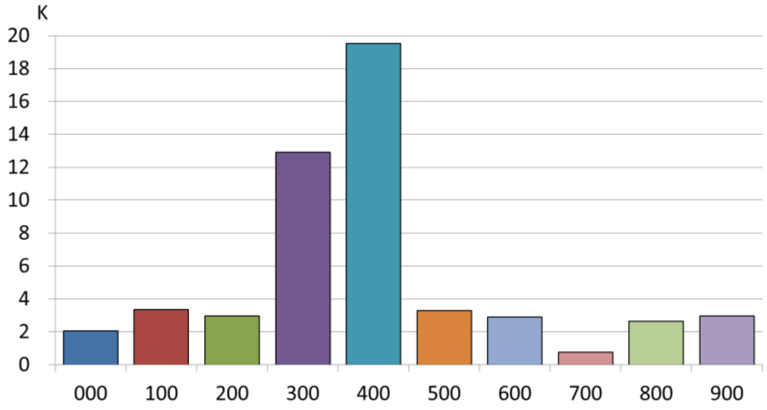

Figure 1. Histogram of the Number of Borrowed Books for NDC Main Categories

Figure 1 shows the distribution of the number of borrowed books (and other materials) for NDC categories from the loan records of Kyushu University Main Library. We can see easily that the most popular main category is 400 (Natural Sciences) and 300 (Social Sciences). This is a reasonable result because 400 is the representative subject field for science and technology and 300 is for social and human sciences.

Another possible explanation why the books in 300 and 400 are so high is that the main building of the Faculty of Sciences (SC) locates next to the Central Library building and the main building for the Faculty of Agriculture (AG) locates at the opposite side of the library. Thus the students in these faculties have a great advantage in visiting and borrowing books than the students in other faculties, and as a result the number of books in the main category of 400 should notably increase.

Actually, the campus for the social and humanity sciences, i.e. Economics (EC), Law (LA), ED (Education) and Letter (LT), locates next to the main campus of Kyushu University, and the Central Library locates near the closest end to the social and humanity sciences campus. Thus this situation should increase the number of borrowed books in the NDC category 300 for social sciences.

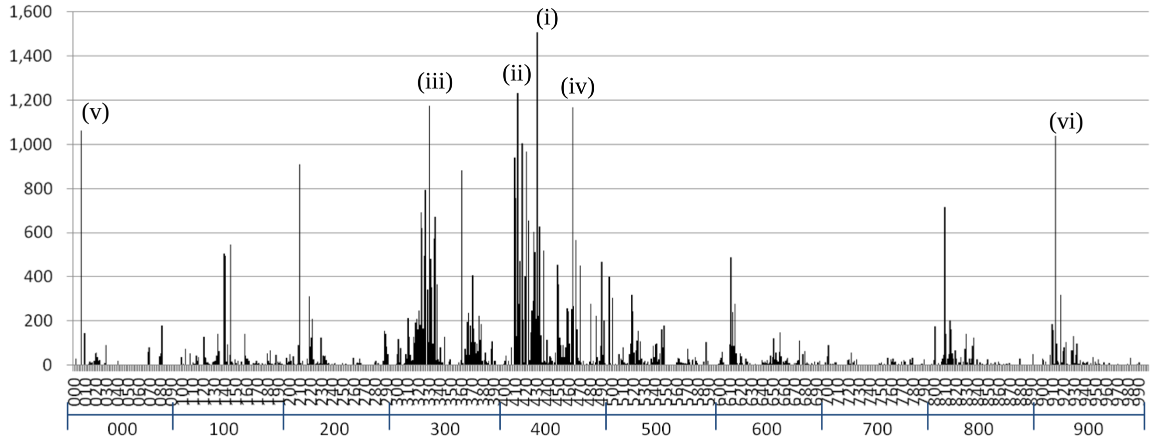

Figure 2. Number of Borrowed Books per NDC Category

In order to investigate further on the loan records themselves, we would have a closer look of the data. Figure 2 shows the distribution of the number of borrowed books, or in the loan records, against the 1,000 NDC categories; from 000 to 999. The maximum value (i) is 1,508 for the category 431 (Physical and Theoretical Chemistry) in the 430 (Chemistry) division in the 400 (Natural Sciences) main category. The next many (ii) is 1,233 for 413 (Analysis) in 410 (Mathematics), which is also in 400 (Natural Sciences).

The 3rd most (iii) is 1,175 for 331 (Economic Theory and Thought) in 330 (Economics) in 300 (Social Sciences), followed by (iv) 1,166 for 464 (Biochemistry) in 460 (Biology) in 400 (Natural Sciences), followed by (v) 1,064 for 007 (Computer Science) in 000 (General Works), followed by (vi) 1,040 for 913 (Japanese Novels) in 910 (Japanese Literature) in 900 (Literature), and so on.

The category 007 for Computer/Information Science is a typical case for showing that NDC, or decimal classification systems themselves, is not a good system for adapting the change of importance of subjects as time goes. In this case, computer science was not in consideration when NDC system was set at the first time. Then this subject field had developed vigorously and when they wanted to assign a category code, there were no available code number left for it. As the result, computer science was squeezed to the category number of 007, which belongs to the category of 000, the general works. Since computer science has the NDC category number in such a low level, the detailed categories in computer science have to go even to after the decimal point; such as 007.1, 007.11, and even the code of 007.304 really appear in our loan records.

3.2 Indexes for Measuring Closeness between NDC Categories

In this section, we define a measuring index between two NDC categories for their closeness relationship at first. Then we extend the definition to compare a section category such as 007 and a main category such as 300. Using this definition, we will investigate the section category of 007 in detail in Section 3.3.

Let LR be the collection of loan records, and let each member $r \in LR$ be a loan record consisting of the book ID, book’s NDC number, patron ID, patron’s type/affiliation information. We use PID(r) and NDC(r) for representing the patron ID and NDC category number in section level for the record r. Let us use NDC as the set of all NDC categories from 000 to 999. Also let M be the collection of the main categories of NDC; i.e. M={000, 100, 200, …, 900}, and let Pat be the set of all patrons who appear in LR. In this study #LR=53,329 and #Pat=6,173. Note that the symbol # indicates the number of the collection.

For $p \in Pat$ and $c \in NDC$, we define the number of borrowed books by the patron p from the NDC category c as follows:

$$nb(p, c) = \#\{r \in LR \mid PID(r)=p, NDC(r)=c\}$$

Thus nb(p, c) indicates the number of the loan records which have p as the borrower and c as the NDC category of the borrowed book.

Then we define the number of patrons who borrowed books from two NDC categories c1 and c2 ($c1, c2 \in NDC$) as the index for the closeness of co-occurrence relationship between c1 and c2.

$$np(c1, c2)=\#\{p \in Pat \mid nb(p, c1)>0, nb(p, c2)>0\}$$

Thus np(c1, c2) indicates the number of the patrons who borrowed at least one book from the NDC category c1 and borrowed also at least one book from the NDC category c2.

From the definition we can see that np is symmetric; np(c1, c2) = np(c2, c1) for $c1, c2 \in NDC$. Also we can see that np(c, c) for $c \in NDC$ is the number of patrons who borrowed at least one book of the NDC category c.

Let us define M(c) for $c \in NDC$ as the main NDC category of c. In formula we can also define it as follows:

$$M(c)=Int(c/100) \times 100,$$ where Int takes the integer part, or floor of the non-negative real number.

For example, M(007)=000, M(331)=300, and M(417)=400.

For $c \in NDC$ and $m \in M$, we define their closeness index by using the number np:

$$npM(c, m)=\sum_{M(c')=m} np(c, c')$$

Alternatively we can define as follows:

$$npM(c, m)=\sum_{c' \in \{m, m+1, m+2, \ldots, m+99\}} np(c, c')$$

From the definition, npM(c, m) for the NDC category c and for the main category m is the sum of the patrons who borrowed books from the category c and from one of the NDC categories which belong to the main category m.

3.3 Investigation of the Category of 007 for Information/Computer Science

In this section, we would like to investigate how the books in category 007 are borrowed; especially in comparison with its closeness to other NDC categories.

Figure 3 shows the npM(007, m) for m from 000 to 900. We can see that the main category of 400 (Natural Sciences) takes the largest number (966). As we have pointed out that 007 is the category for Computer Science and thus it is easy to predict it has high relationship to the main category of Natural Sciences. It is our surprise that the relationship to 300 (Social Sciences) is higher than to 500 (Technology and Engineering). We predicted that the subject field of computer science is a part of engineering and the quite a lot of students who study computer science belong to the faculty of engineering, and thus the students will borrow the books of category 500 (Technology/Engineering) as many as those of 400 (Natural Sciences). We need to investigate further on this topic in order to find out the reason.

Figure 4 shows the ratios of faculties of borrower for the books having NDC category of 007. Against our prediction, the highest share goes to SC (Sciences) instead of TE (Engineering), which we predict the maximum. Thus the books for computer/information science are rather for the science students than those of engineering. At the same time, the ratio of TE is the second highest and higher than the double of the third share, which goes to EC (Economics). The fact that the 3rd highest is EC in interesting; which presumably because the EC students recognize the importance of information and communication technology (ICT) in terms of economical point of view.

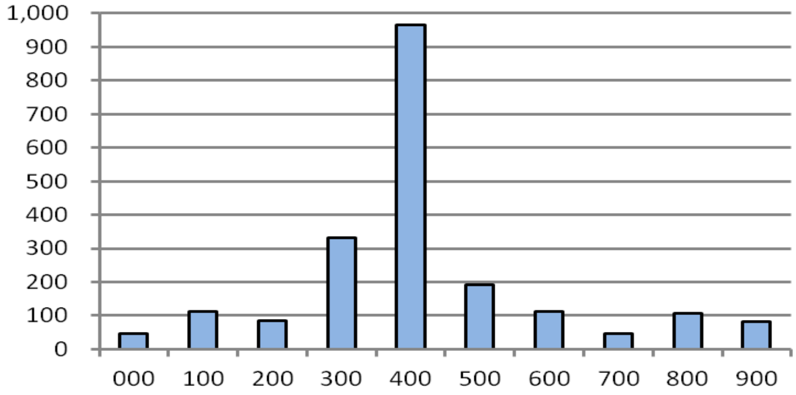

Figure 3. Numbers of Patrons who have the Main NDC Categories Related to the Category 007 (Information Science)

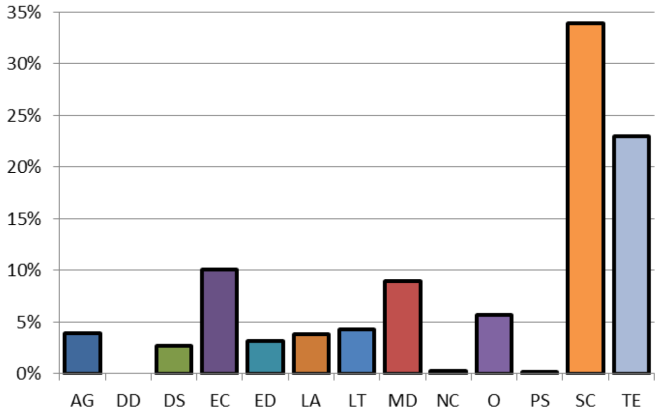

Figure 4. The Borrower Affiliation Rates for the Books of NDC Category 007 (Information Science)

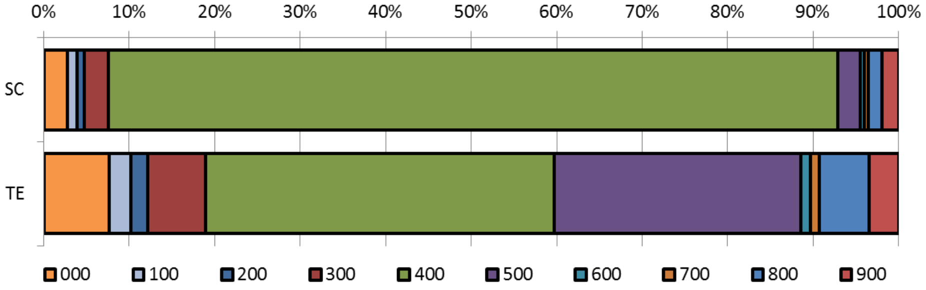

Figure 5. The Interest Area Profiles of Faculties SC (Sciences) and TE (Engineering)

Our next interest is to investigate how much the computer/information science attracts the students of engineering. Figure 5 shows the interest area profiles of SC and TE, in order to compare these two faculties on their interest to subject fields including to computer science.

We can see in Figure 5 that even with that they share NDC 400 (Natural Sciences) as the highest subject field, the ratios are crucially different; 85% for SC and 41% for TE. The 2nd highest NDC category for TE students is 500 (Technology/Engineering) and its share is 29%; whereas it is 3% for SC. Thus the interest to the NDC subject category 500 clearly discriminates the TE students from those of other faculties including SC.

At the same time, the interests to the NDC category 000 (General Works) also differentiate the students of TE from SC. Actually more than 90% of books of NDC 000 in the loan records are 007, we can say that the share 8% of NDC 000 for TE is the share to NDC 007 (Computer/Information Science).

From these investigations, we find that even with relatively smaller interest ratio (3%) to NDC 007 for SC in comparison with that (8%) for TE, the share of books borrowed by SC (20%) is much higher than that of TE (13%), and SC students borrowed quite a lot of books from 400, the most related NDC main category of NDC 007 (Computer/Information Science) becomes 400 (Natural Sciences); which explains why the NDC main category 400 takes the top in Figure 3.

4. CONCEPTION OF VIRTUAL NDC MAIN CATEGORY AND ITS APPLICATION TO INTEREST AREA PROFILING

In this section, we define the concept of the virtual NDC main category as an application of similarity measure we have discussed in Section 3. Then we recalculate the interest area profiles of the 12 faculties plus all the other patrons affiliated in other divisions, and compare the profiles by using the original and the new categories.

4.1 Virtual NDC Main Category

Using the closeness measure npM between a NDC category c and a main NDC category, we define the concept of the virtual main category of c as the main category that is the closest to c in terms of npM values. For c∈NDC, its virtual main category vM(c) is defined as the m∈M where npM(c, m) has the maximum value among npM(c, m') (m'∈M). Thus the following equation holds:

$$npM(c, vM(c)) = \max_{m' \in M} npM(c, m').$$

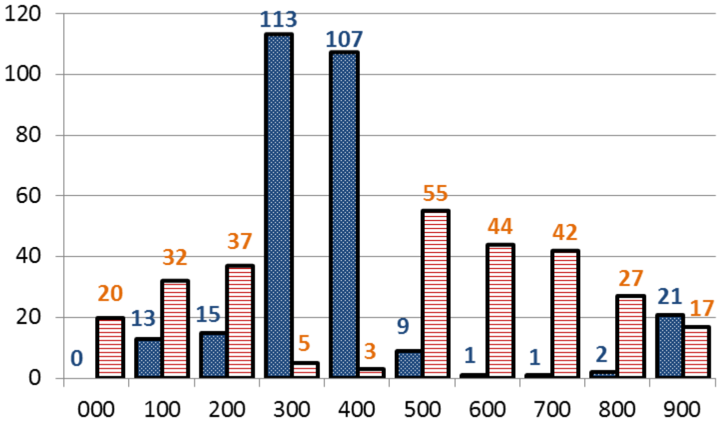

Figure 6. Numbers of NDC Categories Moving to Different Main NDC Category (left: Coming, right: Going)

We are interested in the categories where the virtual main category is different from the original main category; i.e. vM(c)≠M(c). Figure 6 shows the numbers of NDC categories moving into another main category as its virtual NDC (main) category. The left (blue) bars indicate those coming from other main categories, whereas the right (red) bars indicate those going out to other main categories; i.e. for m∈M,

left: #{c∈NDC | vM(c)=m, M(c)≠m}

right: #{c∈NDC | M(c)=m, vM(c)≠m}

We can see that the main category 300 (Social Sciences) takes the top share, followed by 400 (Natural Sciences). Their frequencies are very close each other and the rest main categories are very small in numbers as we compare them with these two tops.

It is easy to see that most moving categories go to either 300 (Social Sciences) or 400 (Natural Sciences) as their virtual main category. These categories come from the categories 500 (Technology/Engineering), 600 (Industry and Commerce), 700 (Arts), 200 (History/Geography), 100 (Philosophy and Religion), and so on, in the decreasing order of the number of providing categories.

As we investigate further, we can see that the main category 000 (General Works) provides 20 categories out of 100 section categories (20%) and has no accepting categories (0%). Considering that the categories ending by the number 0 indicate it is for general topics, or for those books that are difficult to classify into specific topics of the categories/subjects/fields. Thus we can say that use of the virtual category clarifies the not-sufficiently-classified categories so that it becomes easier to recognize the practically related subject categories.

The main category of 900 (Literature) is also the one that have more accepting categories, i.e. 21, than providing categories, i.e. 17. Roughly speaking, about 1 out of 5 (20%) categories move to other main categories and at the same time almost the same amount of categories come and join this category of 900 (Literature). The category 900 is the only one category that the coming and going out numbers are balanced and thus this fact characterizes the category 900 for literature. We need to investigate further what makes this unique fact to the category 900.

The category 500 (Technology/Engineering) is unique in the sense its out-going category number 55 is the biggest among 10 main categories. Also, it becomes the best of the losing number of categories; i.e. 46 by subtracting 9 from 55. Roughly, half of the categories that used to belong to the category 500 move to other categories. In other word, the remaining half categories in 500 are those “absolutely” belong to 500 of Technology/Engineering.

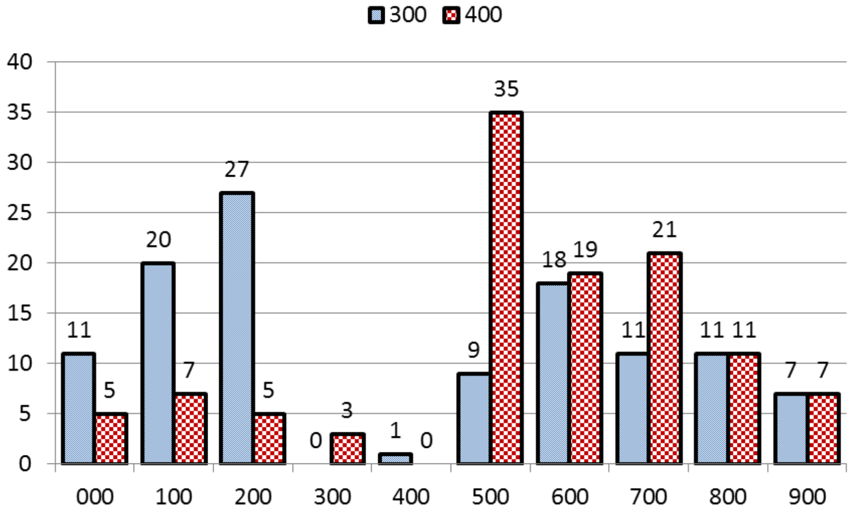

Figure 7. Distribution of Original Top/Main Categories to Categories 300 and 400 as Virtual Categories

Now we know that the main categories of 300 (Social Sciences) and 400 (Natural Sciences) receive most moving categories from other main categories. We would like to investigate further on these two categories. Figure 7 shows the distribution of the numbers of NDC categories' original main categories that have 300 and 400 as their (different) virtual main category, respectively. Thus the numbers at 300 for 300 and at 400 for 400 are 0 because we avoid considering the remaining section categories.

For the categories that have 300 (Social Sciences) as their virtual category, their original main categories are, in the decreasing order of the number of the providing categories, 200 (History and Geography) with 27, 100 (Philosophy and Religion) with 20, 600 (Industry) with 18, and so on. More specifically, among 27 categories coming from the main category 200 (History and Geography), 6 are from the division 20, or 20*, i.e. 200 to 209, 2 from 21, 1 from 22, 5 from 23, 3 from 24, 2 from 25, 2 from 26, 2 from 27, 3 from 28, and 1 from 29. For the division 20, the section categories 200 (History in General), 203 (Reference Book), 205 (Periodicals), 207 (Research/Instruction/Education Methods), 208 (Collections), and 209 (World/Cultural History) have 300, or 3, as their virtual main category.

For the category 400 (Natural Sciences), the maximum number of 35 comes from 500 (Technology); 3 for the division 500 (Technology and Engineering), 3 for 510 (Civil Engineering), 1 for 520 (Architecture), 5 for 530 (Mechanical/Nuclear Engineering), 4 for 540 (Electrical/Electronic Engineering), 2 for 550 (Maritime/Naval Engineering), 4 for 560 (Metal/Mining Engineering), 10 for 570 (Chemical Technology), 1 for 580 (Manufacturing), and 1 for 590 (Domestic Arts and Sciences).

It is interesting to see that all categories in the division 570 for Chemical Technology have 400 for Natural Sciences as their virtual main category. Thus we can see that Chemistry-related subjects are strongly related each other of the natural science and technological/engineering chemistry, i.e. from theory to practice. It is also interesting the second largest is 700 (Arts). Among 700, 4 out of 10 categories in the division 750 (Craft) and 3 in 740 (Photography and Printing) have 400 as their virtual category.

It is also interesting to see that 16 categories in the main category 000 (General Works) move to 300 and 400. Among them, 11 categories move to 300 and just 5 categories move to 400, which is against our prediction because most (more than 90%) borrowed books that belong to 000 are of 007 (Information/Computer Science) books. Actually, the category 007 moves to the category of 400 as its virtual main category. However other, seemingly those of smaller number of books belong to, categories such as 002 (Knowledge/Learning/Academic), 030 (Encyclopedia), 070 (Journalism/Newspapers) move to the main category 300.

4.2 Comparison of Interest Area Profiles of Using Original and Virtual NDC Main Categories

We define the interest area profile as a 10 dimensional vector by the element corresponds to the top-level (main) category of NDC [10-12]. The value of each element of the vector is the ratio of the number of borrowed books from the corresponding main category of NDC. It is defined formally as follows:

Let p be a patron. Then the profile of p is defined by Prof(p) = <r_{000}, r_{100}, …, r_{900}>, where

$r_i = f_i / \sum_{j=000}^{900} f_j$ and $f_i = \#\{ l \in LR \mid \mathrm{PID}(l) = p, \mathrm{M}(\mathrm{NDC}(l)) = i \}$ for i = 000, 100, …, 900.

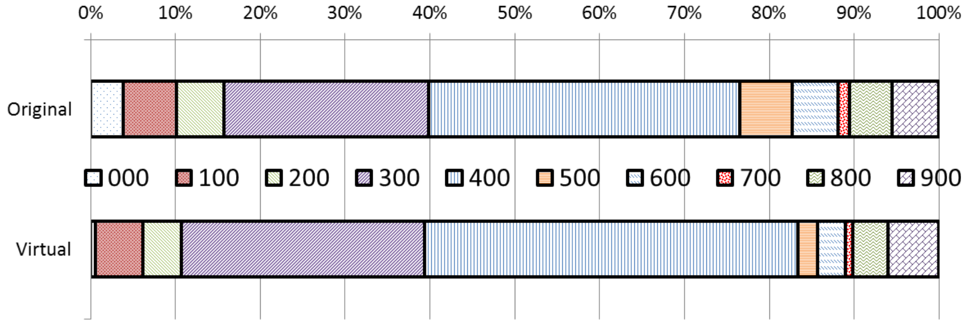

Figure 8. Interest Area Profiles of All Patrons with Original NDC (upper) and with Virtual NDC (lower)

Figure 8 shows the profiles of interest area in order to compare the case of using the original main categories (upper) and the case of using the virtual main categories (lower). More formally the original profile uses the ratios of the numbers #{r∈LR | M(NDC(r))=m} for m=000, 100, …, 900, and the modified profile uses the numbers #{r ∈ LR | vM(NDC(r))=m}; i.e. to use vM instead of M.

We can see in Figure 8 that the categories 300 (Social Sciences) and 400 (Natural Sciences) have more shares than the original ones as we can predict easily from Figure 6. On the other hand the main categories 000 (General Works), 500 (Technology and Engineering), 600 (Industry and Commerce) lose their shares.

For the main category 000 (General Works), 20 categories have virtual main categories different from 000; 11 for 300 (Social Sciences), 5 for 400 (Natural Sciences), 2 for 900 (Literature), and 1 for 100 (Philosophy) and 200 (History). Thus about half of them move to 300, 1/4 to 400, and the rest 1/4 to other categories.

For the main category 500 (Technology and Engineering), among 55 categories, 400 (Natural Sciences) takes the most of 34 categories, followed by 9 for 900 (Literature), 8 for 300 (Social Sciences), 2 for 100 (Philosophy), and 1 for 600 (Industry and Commerce). Thus the main category 500 strongly relates to Natural sciences as the subject field.

For the main category 600 (Industry and Commerce), among 44 categories, 300 (Social Sciences) takes the 18 categories, followed by 16 for 400 (Natural Sciences), 4 for 200 (History), 2 for 500 (Technology and Engineering), and 1 for 800 (Language) and 100 (Philosophy). So in this case the main categories 300 and 400 take the most of the section categories from the main category of 500, which means that industry and commerce fields are related both to natural sciences and technology/engineering field as well as to social sciences, presumably in terms of business matters and something like that.

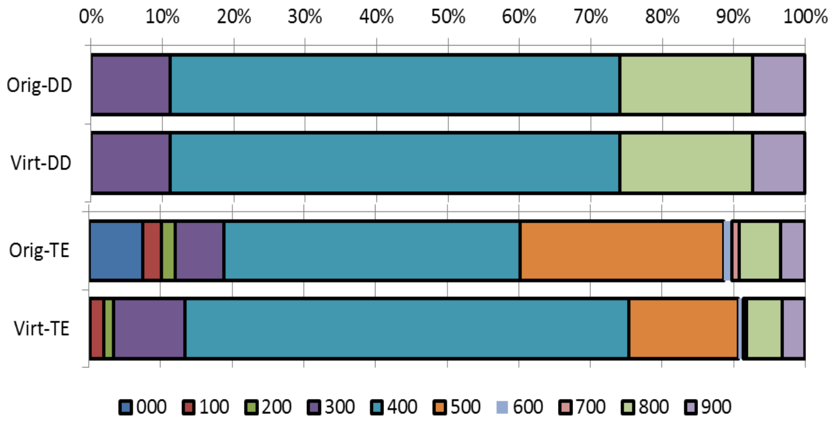

Figure 9. Comparison of Interest Area Profiles of Original and Virtual NDC Main Categories for the Typical 2 Faculties; Having No Difference (DD, upper) and that Having the Maximum Difference (TE, lower) in terms of Cosine Similarity.

Figure 9 shows the comparative interest are profiles of two faculties; the Faculty of Dental Science (DD) and the Faculty of Engineering (TE). In each faculty, the upper bar graph shows the interest area profile using the original NDC categories, whereas the lower one shows the profile using the virtual NDC categories. These faculties are two extreme cases in the sense DD is the example that the two profiles are exactly the same and thus the cosine similarity of the two profiles is 1, whereas TE is the example that gives the most difference (cosine similarity = 0.9266) among the 13 faculties.

As we have seen in Figure 6, the NDC main category 500 (Technology/Engineering) is the major provider of NDC categories thus the share of 500 for TE becomes about half in the Virt-TE graph from the Orig-TE graph. As a result, the share of 400 increases suitably. As we compare the DD and TE in virtual profiles, they are similar in the categories 400 (Natural Sciences) and 300 (Social Sciences). On the other hand, they are quite different in 500 (Technology/Engineering) and 800 (Languages); in TE, 500 takes the 2nd share and in DD, 800 takes the similar share of 500 of TE, which differentiate the faculties DD and TE. It is understandable that the books of “true” Technology/Engineering take the reasonable amount of share in their interest area. On the other hand, it is a kind of surprise that 800 (Languages) books take the similar share for DD patrons. We have to investigate further in order to find why.

5. CONCLUDING REMARKS

In this paper we proposed an index for measuring closeness between an NDC category and a main, or top-level, NDC category and propose a concept of virtual (main) category based on the closeness of them. By applying these to the loan records from Kyushu University in Japan, we found that a lot of categories, especially in the main categories 500 (Technology and Engineering), 600 (Industry and Commerce), 200 (History and Geography), 700 (Arts), etc.

have either 300 (Social Sciences) and 400 (Natural Sciences) as their virtual categories. So we can say that the Social and Natural Sciences dominate the subject field for university patrons.

We also applied the result compare the ratios of borrowed books in two cases; one by using the original main NDC categories and the other by using the virtual NDC categories. As the result, we can see that the ratio for 400 (Natural Sciences) has increased from 37% to 44% (+7%) and for 300 (Social Sciences) it has changed from 24% to 29% (+5%). On the other hand, the main categories 000 (General Works) decreases from 4% to 1% (-3%), and 500 (Technology and Engineering), from 6% to 2% (-4%), and 600 (Industry and Commerce), from 5% to 3% (-2%), also decrease notably.

Then we compared the interest area profiles using virtual NDC and original NDC of 13 faculties; to be exact, 12 faculties plus 1 group of patrons with other affiliations. As we investigated the amount of differences using cosine similarity, the maximum similar value is 1, i.e. exactly the same, for the Faculty of Dental Science (DD), whereas the minimum similarity value is 0.9266 for the Faculty of Engineering (TE). One of the most influencing reasons on this differences between faculties might be that the main category 500 (Technology/Engineering) is the one that provides most section categories to other fields, so that its share in the profile changes most as we compare the profiles between the one using original NDC and the virtual NDC. The faculty of Engineering (TE) is exactly the faculty that its students borrow a lot of books from the main category 500.

Such analysis results may be utilized, for example, in rearrangement of the shelves; from normal arrangement where the shelves are arranged according to the (original) NDC numbers to move some books from the original place/shelves to the special shelves so that these shelves are next to the shelves for the virtual main categories.

The eventual goal of our study is to develop a variety of seeds for e-library services together with of e-library services as their applications so that to increase the patron satisfaction, and thus also to increase the reputation of libraries. In order to advance more steps toward such a goal, we have to study further topics including:

(1) To develop more sophisticated methods for profiling patrons and books

In this paper we use the number of patrons who borrowed books from both NDC categories in order to measure the closeness of two categories. The profiling method of using vectors of frequency, or weight in general, called vector space model (VSM) [15] has been widely studied and used in the field of document retrieval, analysis etc. The study presented in this paper is the beginning trial of using such method to the characterization of the subject fields using decimal classification (DC) system. We have to investigate further on this approach in order to customize this method so that it is more applicable to the subject field classification, understanding, retrieval, and others. Also it will be preferable to use other measures that reflect other aspects such as the importance of the books for the borrowing patrons, the aim of the borrowing in consideration of the attributes of the books, etc.

(2) To utilize other types of data and analyze such heterogeneous data

The candidates of data for analysis include those obtained from library and those from other types of data such as lecture data (such as [4, 9]) in universities and other academic organizations, bibliographic data, purchase data from businesses, etc. The data from libraries include the access log data of OPAC (Online Public Access Catalog) [14] and of institutional repository as well as home pages, patrons' entrance/exit records, seat occupation data, etc.

(3) To perform the feasibility studies for practical library services

Last but not least, we have to put the ideas in technology and in patron services into practice. In order to choose the practically useful ideas, we need to carry out the feasibility studies.

The current status of our study is still in the beginning stage because we have just started taking this approach of this direction, and thus we have a long way to go toward the practical level. However as we consider the potential importance of the studies in this approach, the outcomes will become very important for the library services in the, hopefully 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, 2014, and also partly supported by the Kyushu University Interdisciplinary Programs in Education and Projects in Research Development (P&P).

REFERENCES

[1] Cohen, S., Fereira, J., Horne, A., Kibbee, B., Mistlebauer, H., and Smith, A., 2000. MyLibrary Personalized Electronic Services in the Cornell University Library. D-Lib Magazine, 6(4).

http://www.dlib.org/dlib/april00/mistlebauer/04mistlebauer.html.

[2] 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.

[3] Di Giacomo, M., Mahoney, D., Bollen, J., Monroy-Hernandez, A. and Ruiz-Meraz, C.M., 2001. MyLibrary, A Personalized Service for Digital Library Environments. Proc. of the 2nd DELOS Workshop on Personalization and Recommender Systems in Digital Library. 6pp.

[4] Goda, K., Hirokawa, S., and Mine, T., 2013. Correlation of Grade Prediction Performance and Validity of Self-Evaluation Comments, Proc. ACM SIGITE2013. pp. 35-42.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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, Portugal. 8pp.

[12] Minami, T. and Baba, K., 2014. Knowledge Field Reorganization for the Library Patrons' Interest Area Analysis – An Investigation for Next-Generation e-Library Services, 7th IADIS International Conference on Information Systems 2014 (IS 2014), Madrid Spain, pp. 103-110.

[13] National Diet Library. 2003. Bibliographic Data Creation Tool. http://www.ndl.go.jp/jp/library/data/zan9-old.html (in Japanese)

[14] Online Computer Library Center, Inc. (OCLC). WorldCat Collection Analysis.

http://www.oclc.org/collectionanalysis/

[15] Salton, G., Wong, A., and Yang, C.S., 1975. A Vector Space Model for Automatic Indexing. CACM, Vol. 18, No. 11. pp.613-620. http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf