Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2013) 000–000

Procedia
Social and Behavioral Sciences

www.elsevier.com/locate/procedia

3rd International Conference on Integrated Information (IC-ININFO)

Automatic Classification of Reference Service Records

Shunsuke Arai[a]* and Keita Tsuji[b]

[a]*Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2 Kasuga, Tsukuba-city, Ibaraki-ken 305-8550, Japan*
[b]*Faculty of Library, Information and Media Science, University of Tsukuba, 1-2 Kasuga, Tsukuba-city, Ibaraki-ken 305-8550, Japan*

Abstract

The National Diet Library in Japan maintains a database of reference service questions and the answers given to them (henceforth, reference service records). The questions are submitted to public and university libraries in Japan by users, and the answers are given by the libraries. To improve the findability of these records, we propose a method for automatically assigning Nippon Decimal Classification (NDC) codes to them. Although some studies have been conducted on classification of the reference service records, their precision is unsatisfactory. One reason for this is that these methods depend only on the texts of the reference service records. Many reference service records contain book titles that the reference librarians believe will be helpful to the questioners. The NDC codes of these books might be useful for classification. Based on this background, we propose to automatically (1) extract the book titles in the reference service records, (2) submit them to the NDL-OPAC, (3) obtain the NDC codes of the books, and (4) input them into a support vector machine (SVM) as features together with the texts of the reference service records, and then classify them. A total of 62,884 records were used. For the first digits of the NDC codes, our method achieved a 53.4% precision. For the first and second digits of the NDC codes (i.e., classes and divisions), our method achieved a 46.1% precision. These figures are significantly higher than those of the preceding studies.

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.

Keywords: reference service records; automatic classification; NDC codes;

* Corresponding author. Tel.:+81-46-848-8518;
E-mail address: syun0201@gmail.com

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.

1. Introduction

The National Diet Library in Japan maintains a database of reference service records, called the "Collaborative Reference Database." Each reference service record consists of reference questions and the answers given to them. The questions are submitted to public and university libraries in Japan by users, and the answers are given by the libraries. According to the "2012 Business Report on the Collaborative Reference Database Project," a total of 583 institutions including public and university libraries are participating in the Collaborative Reference Database. A total of 62,884 reference service records were contained in the database as of July 2013.

In this database, the reference service records are described in accordance with a specified format. As shown in Table 1, the use of "keywords," "type of subject," and "Nippon Decimal Classification (NDC) codes" is optional, and the NDC codes are assigned to only 40,288 records because such codes can be burdensome to reference librarians. However, NDC codes allow users to easily find reference service records and are therefore useful. Based on this background, we propose a method for automatically assigning NDC codes to these records.

The reference materials (such as dictionaries, bibliographies, and related books) indicated in Table 1 are those used to answer the reference questions. These materials often have NDC codes of their own. We assume that their NDC codes and the codes that should be assigned to the reference service records are closely related, and therefore propose using the former to determine or assign the latter.

Table 1. Format used to complete the reference service records

No	Item name	Category
1	Question	Mandatory
2	Publication level	Mandatory
3	Control Number	Mandatory
4	Answer	Mandatory
5	Creation date	Optional
6	Resolved/Unresolved	Optional
7	Keyword	Optional
8	NDC codes version	Optional
9	NDC codes	Optional
10	Research type	Optional
11	Type of subject	Optional
12	Reference materials	Optional
13	Answering process	Optional
14	Reference	Optional
15	Case research matter	Optional
16	Note	Optional
17	Inquirer category	Optional
18	Contributor	Optional
19	Related image	Optional
20	Registration Number	Automatic assignment
21	Registration date/time	Automatic assignment
22	Last update date/time	Automatic assignment
23	Offering library code	Automatic assignment

2. Previous studies

Yoda (2006) pointed out that participation in the Collaborative Reference Database can increase the burden placed on staff members, and attempts have been made to improve the operational flow for a registration of the reference service records. Tsutsumi et al. (2011) pointed out that the reference service records in the Collaborative

Reference Database have been recognized as useful throughout the Web. It will therefore be useful to automatically assign NDC codes to reference service records (1) to reduce the costs for librarians who register the records, and (2) to enhance their findability or searchability.

Experiments on automatically assigning NDC codes to the reference service records in the Collaborative Reference Database were conducted by Harada et al. (2007). Using a support vector machine (SVM), the authors tried to assign the first digits of the NDC codes (class), and the first and second digits of NDC codes (sub-division), to the records (similar to Dewey Decimal Classification, NDC codes mainly consist of three-digit numbers, for instance, 324 represents "international law," 32* represents "law" (as a sub-division), and 3** represents "social sciences" (as the class)). For the learning samples, 6,337 reference service records with two- or three-digit NDC codes were used. As a result, the precision and recall rates for the first digit assignment were 46.5% and 38.4%, respectively. Harada et al. (2007) pointed out that many errors occurred because of the irrelevant words contained in the records.

Experiments on automatically assigning NDC codes to books rather than reference service records include an experiment by Ishida (1998), who used 38,011 book titles and their NDC codes for machine learning. The precision of the NDC code assignment was 55.9%. Ishida (1998) pointed out that using only the meaningful parts of the titles significantly enhanced the precision. A similar attempt was made by Agata et al. (1999) who automatically assigned NDC codes to Web pages.

3. Automatic assignment of NDC codes to reference service records

A method for automatically assigning NDC codes to reference service records is explained in this Section. We propose the use of reference materials to automatically assign NDC codes to reference service records. As we previously mentioned, it is likely that the NDC codes of the reference materials are related to the NDC codes that should be assigned to the reference service records. The NDC codes of the reference materials can be a good clue for machine learning (such as an SVM) to determine the NDC codes of reference service records. We therefore propose assigning NDC codes to the records as follows:

(1) Obtain the reference service records registered in the Collaborative Reference Database.
(2) Extract the titles of the reference materials in the records.
(3) Conduct an *NDL Search* using OpenSearch[†] to extract the NDC codes of the titles extracted in (2).
(4) Use the NDC codes obtained in (3) and other clues to automatically assign NDC codes to the records through machine learning. As described later, Weka[‡] 3.6.6. (free software for machine learning) was used to achieve this.

3.1. Extracting the titles of the reference materials

In the Collaborative Reference Database, the bibliographic information of the reference materials is written in the designated places in each reference service record. However, they are not consistent partly because these reference service records are written by various staff workers at various libraries. Examples of bibliographies found in the records are as follows.

- 1.『Exploring Architecture in Chiba』 (Tetsuo Nakamura, Ronsyobou 2004) | 0200797764.
- 【Material 1】「A Dictionary of Composers and their Works in Classic Music」 (Sanseido, 2009).

[†] http://www.opensearch.org/Home
[‡] http://www.cs.waikato.ac.nz/ml/weka/

- Encyclopedia of Japanese Myths / Compiled by Shuhei Aoki [, etc.] Daiwa Shobo, 1997 ISBN: 4479840435 p.31.

Extracting the titles of the reference materials from these records is therefore not an easy task. To do so, we used the following heuristics: (1) Extract the character strings enclosed in the Japanese brackets 『』. (2) If 『』 and "http" character strings are not used, and the character strings are separated with a space and comma, the first character strings are extracted as titles, excluding noise such as 【Material 1】.

3.2. Identifying the NDC codes of the reference materials

After extracting the titles of the reference materials, we input them into an *NDL Search* using OpenSearch and identified the NDC codes[§]. *NDL Search*, provided by the National Diet Library, offers a bibliography of books, papers, and other information. We can obtain the NDC codes of books as a part of their bibliography.

Since the titles extracted in Section 3.1 may not be perfect, we decomposed them into morphemes using MeCab[**], a Japanese morphological analyzer, and input the nouns into an *NDL Search*. For instance, for the title, *Exploration of Architecture in Chiba*, we input "Chiba, architecture, exploration" and ran a search. The search results include bibliographies of one or more books containing "Chiba," "architecture," and "exploration" in the title. When two or more books were hit, we decomposed their titles into morphemes in the same way and determined the target reference materials. Book *i*, which had the highest similarity with S_i, which is defined as follows, was regarded as the target reference material.

$$S_i = \frac{2n(N_R \cap N_i)}{n(N_R) + n(N_i)}, \tag{1}$$

where N_R and N_i represent the sets of nouns in the title of the reference material and book *i*, respectively, and $n(N)$ represents the number of elements in set *N*. The NDC code of book *i* is extracted from the bibliography and regarded as the NDC code of the reference material.

3.3. Determining the NDC codes for the reference records

There are ten different first digits that can be used for an NDC (ranging from 0 to 9). Based on this, we represent the reference service records using 10-dimensional vectors. The n-th element of the vector represents the number of reference service materials whose first digit of the NDC contained in the records is "n-1." For instance, suppose a case in which four reference materials are contained in one reference service record, and their NDC codes are 000, 102, 103, and 110. The first element of the vector is 1 (corresponding to an NDC of 000), the second element of the vector is 3 (corresponding to NDCs of 102, 103, and 110). The value of the other elements is 0.

As described later, some reference service records are already assigned an NDC. We used such records (i.e., the pairs of NDCs and the above-mentioned vectors) as learning data, and as testing data for an SVM. For our experiment, we used a three-fold cross validation.

In addition, we conducted an experiment to automatically assign the first and second digits of the NDC. In this case, the vector has 100 dimensions. If we use the above example again, the first element of the vector is 1 (corresponding to an NDC of 000), the ninth element of the vector is 2 (corresponding to NDCs of 102 and 103), and the tenth element is 1 (corresponding to an NDC of 110). The value of the other elements is 0.

[§] http://iss.ndl.go.jp/?locale=en&ar=4e1f
[**] https://code.google.com/p/mecab/

4. Data used in the experiment

We obtained 62,328 reference service records registered in the Collaborative Reference Database using its API (on March 19, 2013). For our experiment, we used 23,332 records that have both NDC codes and reference materials. A total of 32,657 reference materials are described in the records. Identification of their NDC codes using an *NDL Search* was performed during the period of May 10 to May 12, 2013.

5. Results and discussion

The results obtained from the experiment are discussed in this Section.

5.1. Extracting the titles of the reference materials

As mentioned in Section 3.1, the following two types are extracted as titles: (1) character strings enclosed in the Japanese brackets 『』, and (2) the first character strings if 『』 and "http" character strings are not used, and if the character strings are separated with a space and comma.

As for (1), 『』 was included in 15,905 of all 32,657 reference service records. From these, we randomly extracted 500 reference service records and conducted an error analysis. It was found that the title extraction was unsuccessful for the following two records (0.4%).

- "「Major General Kovalevskii Did Not Exist」（『Sankei Shinbun 2010/10/19 3rd page』）"
- "Chinese Classic Medicine and Japan: Bibliography and Tradition / Hiroshi Kosoto, Hanawa Shobo 1996 ISBN: 4827311420 p.346～ Section 3『Shohinho』"

When an identification of the titles was attempted using (2), but not (1), the character strings that were not titles were extracted in 50 out of 500 cases (10%).

For (2), 16,752 records (=32,657-15,905) did not include 『』. Among these, 1,287 records included "http" character strings. We therefore randomly extracted 500 out of 15,465 (=16,752-1,287) records that did not include 『』 or "http" character strings, and conducted an error analysis. It was found that the title extraction was unsuccessful in 17 cases (3.4%), including the following examples.

- "Illustration: History of Narita, Promotion Version, Compiled by the Editorial Committee of the History of Narita, Narita City 1994.12 L210"
- "Compiled by the National Association for the Study of Educational Methods, Encyclopedia of Contemporary Educational Methods, Tosho Bunka, 2004. Described in p.293."
- "Literature survey within the library"

Eleven of the 17 cases did not contain titles, such as the "Literature Survey within the library" described above.

From these results, we can state that titles were successfully extracted in approximately 15,843 cases for (1) and 14,940 cases for (2), i.e., 30,783 out of the total 32,657 cases. The sampling distribution is 0.94 when the book titles can be extracted. When the titles can be extracted, the population rate R is estimated to be $0.925<R<0.955$ with a sample size of n = 1,000 and a reliability of 95%. Based on the above, it seems highly likely that among all the reference service records, the titles of books can be extracted at a ratio of at least 92.5%.

5.2. Identifying the NDC codes of the reference materials

The results of an *NDL Search*, the NDC codes of the reference materials, and the NDC codes of the reference service records can be categorized into the following four cases. (A) A book can be accurately searched, and the NDC codes of the book matches the NDC codes of the reference record. (B) The book is successfully found; however, the NDC codes do not match the NDC codes of the reference record. (C) The book search is unsuccessful, and the NDC codes do not match the NDC codes of the reference record. (D) The book search is unsuccessful; however, the NDC codes accidentally match the NDC codes of the reference record.

To extract the NDC codes, 100 random extractions from the reference materials in the reference service records were conducted, and the results for cases (A), (B), (C), and (D) are shown in Table 2. The experiment was conducted five times to reduce any inherent bias, and only the first and second digits of the NDC were considered.

Table 2. NDC code extraction results from book titles

	A (%)	B (%)	C (%)	D (%)
1st time	48	42	10	0
2nd time	49	46	5	0
3rd time	58	38	4	0
4th time	45	46	9	0
5th time	47	49	4	0

As a result, the average proportions were 49.4%, 44.2%, 6.4%, and 0% for cases (A), (B), (C), and (D), respectively.

Case (A) made up half of the extractions. Case (B) made up nearly the other half of all extractions; however, the NDC codes obtained by case (B) are only used for forming 10- or 100-dimensional vectors (for an SVM) and are not simply assigned to the reference service records. We can expect some of the extractions for case (B) to contribute to the assignment of proper NDC codes.

For cases (C) and (D), very few extractions were made, i.e., 6.4% and 0%, respectively; therefore, the books seem to be accurately found from their titles.

5.3. Determining the NDC codes for the reference records

The results of an automatic assignment of the first digit of an NDC are shown in Table 3. When the first digit of an NDC code in the reference records is 3, we can see in Table 3 that the assignment precision is 40.6% and the recall rate is 69.1%. Precision is defined as the proportion of the correct NDC codes against all the assigned NDC codes as positive example, where positive example is defined as the reference records that were assigned to each NDC code by categorization. Recall is defined as the proportion of the assigined NDC codes as positive example against all the correct NDC codes.

Table 3. Results of an automatic assignment of NDC codes (first digit of the NDC)

NDC codes	Precision (%)	Recall (%)	Number of cases
0	35.5	16.1	2,189
1	51.8	50.2	1,308
2	58.8	54.5	5,300
3	40.6	69.1	4,589
4	59.6	53.4	1,617
5	61.5	44.1	1,729
6	54.6	44.8	1,393
7	64.2	52.2	2,207
8	57.4	55.8	754
9	62.8	55.7	2,246

The weighted average precision in this experiment was 53.4%, and the weighted average recall rate was 52%, where weighted average (precision / recall) is a sum of average (precision / recall) multiplied by the ratio of its case against all the cases. These values are higher than those of Harada et al. (2007), which we mentioned in Section 2. The same NDC codes were successfully assigned as learning samples in approximately half of the cases using NDC codes in the reference materials.

The most frequently observed error was 3 being assigned as the first digit of an NDC code when the actual number was 2. A total of 1,154 erroneous cases belong to this type. In addition, in a total of 732 cases, a 3 was inappropriately assigned instead of a 0. Likewise, in 553 cases, a 3 was assigned when the actual number was 7.

In many cases, the first digits of the NDC codes were erroneously judged as being a 3. This made up approximately 40% of all errors. This seems to be the reason for the high recall when the first digit of the NDC codes was a 3. Based on the above situation, the overall precision seems to improve by weighting in cases in which the first digit of the NDC codes is determined to be a 3.

Next, the results of an automatic assignment of the first and second digits of the NDC codes are shown in Tables 4 and 5. Table 4 shows the ten highest precisions obtained from the experiment, and Table 5 includes the ten most significant influences.

Table 4. Results of automatic assignment of NDC codes (first and second digits of NDC; ten highest precisions)

NDC codes	Precision (%)	Recall (%)	Number of cases
65	88.9	21.6	74
92	81.4	41.2	308
82	80.0	11.1	72
60	77.8	10.3	68
79	75.8	29.4	85
71	75.0	3.4	87
14	73.3	16.4	67
34	70.0	7.4	94
41	66.7	9.1	44
69	66.7	3.2	62

The weighted average precision in this experiment was 45.6%, and the weighted average recall rate was 36.1%. The same NDC codes as in the learning samples were successfully assigned in approximately half of the cases.

The erroneous assignment of an NDC code of 21 occurs frequently, which is the reason for the high recall rate of 21 in Table 5, which is similar to the experiment on the first digit of an NDC. This is considered to be a significant cause of the overall lowered precision.

Table 5. Results of automatic assignment of NDC codes (first and second digits of NDC; top-ten cases)

NDC codes	Precision (%)	Recall (%)	Number of cases
21	13.6	63.0	2,108
28	40.8	39.4	1,601
91	59.8	55.2	1,552
38	51.2	40.7	953
29	37.1	37,8	943
02	28.9	8.4	784
37	56.1	52.2	760
09	3.4	0.1	682
32	60.0	41.8	625
31	46.8	33	621

In an NDC, a 09 code represents “valuable books, local materials, and other special collections.” This category is expected to have less commonality among all the reference materials, which is considered to be the cause of the

very low precision and recall rate despite the fact that the number of cases belonging to code 09 is high. In addition, there is an exhaustive listing of NDC codes including 02, and therefore both the precision and recall rate are considered to be lower for a similar reason.

When the first and second digits of an NDC code are 96 or 85, the total number of which is few, not even one NDC code was accurately determined in 32 cases, which seems to be due to an insufficient number of learning samples. To solve this problem, we can use different clues in the reference service records, e.g., question and answer sentences.

6. Conclusions

We proposed a method for automatically assigning NDC codes to reference service records using the NDC codes of reference materials. This method consists of (1) extracting the titles of the reference materials from the records, (2) finding their NDC codes using an *NDL Search*, and (3) using the codes as learning data for an SVM. We found that the proposed method is more effective than previous methods proposed in related studies.

Future tasks include examining the effectiveness of using (a) other machine learning methods such as random forest and (b) other information sources for the relationship between NDC codes and their corresponding texts, such as the Japanese National Bibliography (which contains the titles and NDC codes of books published in Japan).

References

AGATA, Teru, ISHIDA, Emi, KUNO, Takashi, NOZUE, Michiko, and UEDA, Shuichi (1999), "Automatic Classification of World Wide Web Pages: Classification Scheme with NDC codes and Classification Using Yahoo Categories," Research Report Database of the Information Processing Society of Japan, System Study Group Report, vol.99, no.39, pp.113-120.

HARADA, Takashi, ETO, Masaki, and ONISHI, Minako (2007) "Automatically Classifying the Reference service records into NDC codes," Journal of Japan Society of Information and Knowledge, vol.17, no.2, pp.61-64.

ISHIDA, Emi (1998) "An Experiment of Automatic Classification of Books Using Nippon Decimal Classification," Library and Information Science, vol.39, pp.31-45.

National Diet Library. Japanese National Bibliography, National Diet Library. <http://www.ndl.go.jp/jp/library/data/syoshiservice.html#2>.

TSUTSUMI, Megumi, SATOU Kumiko, and MAKINO, Megumi (2011) "New Horizon of Reference Service Pioneered by CRD (Cooperative Reference Database)," The Journal of Information Science and Technology, vol.61, no.5, pp.187-193.

YODA, Norihisa (2006) "Digital Reference Service on Collaborative Reference Database Project of the National Diet Library," 2006, The Journal of Information Science and Technology, vol.56, no.3, pp.90-95.