Books Cited in Wikipedia: Possibility to Use their Nippon Decimal Classification Categories for Book Recommendation

Keita Tsuji
Faculty of Library, Information and Media Science, University of Tsukuba
1–2 Kasuga, Tsukuba City, Ibaraki-ken, Japan
keita@slis.tsukuba.ac.jp

***Abstract*—This paper investigated the effectiveness of developing a book recommendation system based on books cited in Wikipedia articles, focusing on their Nippon Decimal Classification (NDC). Among 95,194 articles, 28,154 cited books showing ISBNs in their bibliographies. In many cases, all NDCs of books cited in each article were identical and thus consistent. Such articles can be used for automatic assignment of NDCs.**

Keywords—NDC, Nippon Decimal Classification, Wikipedia

I. Introduction

Wikipedia is now an important source of information for university students. It has been shown that they use Google first and then read Wikipedia articles even in university libraries. Actually, students who search for books and read them are relatively few [1]. In this situation, the present researchers are now developing a system that recommends books to users based on Wikipedia articles they are reading in libraries (the system will be added onto web browsers in libraries' desktop PCs). We hope that students reading Wikipedia articles will become interested in these library books and learn deeply by reading them.

Regarding book recommendation, Nippon Decimal Classification (NDC) categories are found to be effective clues for determining which books should be recommended [2][3][4]. More specifically, machine learning methods, such as the Support Vector Machine, can more effectively find relevant books by using NDC book categories of interest to users. Therefore, if we could assign NDC categories to Wikipedia articles, they might be used for effective book recommendations. While some studies have focused on automatically assigning NDC categories, for instance, to reference records in libraries [5], relatively few have focused on Wikipedia articles.

As a first step toward automatic assignment of NDC categories to Wikipedia articles, the present research investigated NDC book categories cited in Wikipedia. We believe that Wikipedia articles fall into the same (or similar) NDC categories as those of books cited in articles. For instance, if article X cites many books whose NDC categories are "324," X is likely to belong to the "civil law" category. On this basis, we can automatically assign NDC categories to Wikipedia articles. However, this idea cannot be used if (A) many articles do not cite books, or (B) each article tends to cite books with various NDC categories. Therefore, the present research investigated whether objections (A) and (B) hold true. Henceforth, the NDC category is referred to as "NDC" for brevity.

II. Method

We obtained and investigated data in the following steps:

(1) We downloaded the dump of Japanese Wikipedia from https://dumps.wikimedia.org/jawiki/ (as of February 21, 2015).

(2) We extracted pages whose namespace (ns) tag was "0," which means that the page was "Main/Article" (henceforth referred to as "articles").

(3) From pages extracted in (2), we extracted pages containing "References."

(4) We extracted pages where "References" contained bibliographies of books showing their ISBNs.

(5) We searched for correct and consistent bibliographies of books by using the ISBNs in *OpenSearch*, provided by the National Diet Library of Japan.

(6) We eliminated bibliographies of books that did not contain triple-digit NDCs (i.e., which consisted of Main Class, Division, and Section) or titles.

(7) We counted the number of NDCs (only Main Class and Division) in all bibliographies obtained through (6).

(8) We counted the number of NDCs (only Main Class and Division) in each Wikipedia article, identifying the most popular NDC among them and counting its frequency in the bibliography. For instance, if an article contained five books and their NDCs were 324, 324, 324, 325 and 367, NDC "32" (Main Class and Division of NDC 324 and 325. Incidentally, 32 representing "Law" and 36 representing "Society") was most popular, and its frequency was four.

III. Results

The Japanese Wikipedia totaled 2,689,050 pages, and the number of articles obtained through Step (2) was 1,104,962. The number of articles obtained through (3) and (4) were 95,194 and 28,154, respectively. The number of ISBNs contained in 28,154 articles, and thus used in step (5), was 66,295 (including duplicates). The number of bibliographies obtained through (7) was 61,173 (including duplicates). The

978-1-4673-8985-3/16 $31.00 © 2016 IEEE
DOI 10.1109/IIAI-AAI.2016.247

number of articles that contained at least one such bibliography was 26,879. Therefore, NDCs were obtained from 28.2% (= 26,879 / 95,194) of articles

Regarding the possible objection (A), 28.2% of articles cited books from which we could obtain NDCs. In the present study, we focused only on books cited with ISBNs, because citations in "References" without ISBNs are (a) sometimes journal articles, or (b) in various formats, which makes it difficult to extract bibliographic items automatically and correctly (for instance, automatically distinguishing titles from author/publisher names), to input them into *OpenSearch*. However, provided such methods are established, a sufficient quantity of books appears to be cited in Wikipedia that could be used for automatic assignment of NDCs to articles.

Results of Steps (7) and (8) are shown in TABLEs I and II, respectively. In TABLE I, for instance, the most frequently observed NDC Main Class is "2" (History), and the combination of Main Class and Division is "21" (Japanese History). Their frequencies are 16,438 and 6,328, respectively.

We can see in TABLE II that frequencies of the most popular NDC in each reference sometimes equal the number of books cited in references. For instance, concerning 2,795 (= 350 + 1,082 + 1,363) references in which three books were cited, in 1,363 references, the frequency of the most popular NDC was also three. In other words, all three NDCs were identical in 1,363 references. However, many references consisted of various NDCs. For instance, again concerning the 2,795 references in which three NDCs were obtained, in 350 references, frequency of the most popular NDC was one. That means three NDCs differed from each other in these 350 references. Therefore, regarding objection (B), we should consider the variety of NDCs in references when using them for automatic assignment of NDCs. However, it is worth noting that 20,722 (= 14,295 + 3,836 + 1,363+…+ 25 + 17 + 7) references consisted of just one NDC, among 26,879 references (77.1%). Therefore, we might be able to use such references for automatic assignment of NDCs without a complicated process. For instance, Wikipedia page "Fermat's Last Theorem" cited 17 books, and all their NDCs (Main Class and Division) were "41" (Mathematics). In such cases, we can simply and safely assign NDC "41" to "Fermat's Last Theorem."

IV. Conclusions

To develop a book recommendation system, we investigated NDCs of books cited in Wikipedia articles. Results indicated that a sufficient number of articles can be used for automatic assignment of NDCs. On the basis of this finding, we will subsequently develop a system that recommends books to Wikipedia readers in conjunction with NDCs.

References

[1] T. Anbiru et al., "Information seeking behavior," Proceedings of the Spring Meeting of the Japan Society of Library and Information Science, 2010, pp. 87–90. (text in Japanese).

[2] K. Tsuji, F. Yoshikane, S. Sato, and H. Itsumura, "Book Recommendation Using Machine Learning Methods Based on Library Loan Records and Bibliographic Information," International Journal of Academic Library and Information Science, vol. 3, no. 1, 2015, pp. 7–23.

[3] K. Tsuji, F. Yoshikane, S. Sato, and H. Itsumura, "Book Recommendation Using Machine Learning Methods Based on Library Loan Records and Bibliographic Information," Proceedings of the 5th International Conference on E-Service and Knowledge Management (ESKM 2014), 2014, pp. 76–79.

[4] K. Tsuji et al., "Book Recommendation based on Library Loan Records and Bibliographic Information," Proceedings of the 3rd International Conference on Integrated Information (IC-ININFO 2013), 2013, 8p. (No Pagination).

[5] S. Arai and K. Tsuji, "Automatically Assigning NDC Categories to Reference Service Records by Using Machine Learning Methods," Journal of Japan Society of Information and Knowledge, vol. 25, no.1, 2015, pp. 23–40. (text in Japanese).

TABLE I. Frequency of NDCs of Books Cited in Wikipedia

NDC	n	NDC	n	NDC	n	NDC	n	NDC	n	NDC	n	NDC	n	NDC	n	NDC	n	NDC	n
0	1,073	1	3,546	2	16,438	3	8,002	4	8,339	5	5,794	6	4,275	7	9,948	8	533	9	3,225
00	422	10	43	20	756	30	497	40	268	50	223	60	36	70	277	80	98	90	175
01	144	11	64	21	6,328	31	1,393	41	712	51	280	61	138	71	29	81	287	91	2,154
02	96	12	286	22	1,217	32	968	42	347	52	395	62	60	72	1,149	82	50	92	99
03	159	13	193	23	1,465	33	617	43	112	53	1,320	63	1	73	9	83	52	93	293
04	32	14	721	24	109	34	20	44	162	54	659	64	1,173	74	158	84	8	94	378
05	23	15	76	25	164	35	16	45	1,435	55	1,703	65	80	75	373	85	0	95	44
06	15	16	650	26	13	36	1,165	46	486	56	76	66	185	76	1,004	86	3	96	14
07	125	17	396	27	17	37	406	47	837	57	140	67	188	77	2,357	87	0	97	9
08	55	18	636	28	3,676	38	1,093	48	2,224	58	199	68	2,270	78	3,786	88	0	98	19
09	2	19	481	29	2,693	39	1,827	49	1,756	59	799	69	144	79	806	89	35	99	40

TABLE II. Number of Books Cited in One Article and Frequency of the Most Popular NDC Shown in them

		Frequency of the Most Popular NDC													
		1	2	3	4	5	6	7	8	9	0	11	12	13	14
Number of Books Cited	1	14,295	–	–	–	–	–	–	–	–	–	–	–	–	–
	2	2,012	3,836	–	–	–	–	–	–	–	–	–	–	–	–
	3	350	1,082	1,363	–	–	–	–	–	–	–	–	–	–	–
	4	79	376	407	564	–	–	–	–	–	–	–	–	–	–
	5	12	166	194	152	274	–	–	–	–	–	–	–	–	–
	6	6	60	113	110	85	141	–	–	–	–	–	–	–	–
	7	1	19	59	55	50	40	72	–	–	–	–	–	–	–
	8	0	4	28	31	27	21	33	41	–	–	–	–	–	–
	9	0	2	15	14	25	13	17	27	41	–	–	–	–	–
	10	0	1	7	9	15	11	14	8	14	23	–	–	–	–
	11	0	1	1	11	10	7	14	5	8	7	23	–	–	–
	12	0	0	1	1	4	7	4	8	6	3	4	25	–	–
	13	0	0	0	4	6	3	2	3	6	4	4	8	17	–
	14	0	0	1	2	2	2	3	2	2	5	1	9	3	7
	15+	219													