Book Recommender System for Wikipedia Article Readers in a University Library

Keita Tsuji
Faculty of Library, Information and Media Science,
University of Tsukuba
Tsukuba-city, Japan
keita@slis.tsukuba.ac.jp

***Abstract*—Wikipedia has emerged as an important source of information for university students. It has been reported that university students tend to read and do research with Wikipedia articles more than they do with books, even when within a library. To encourage students to read library books as a more reliable source of information, a library system was developed for recommending library books to Wikipedia readers within a particular university library. The proposed system assigns a Nippon Decimal Classification (NDC) category to each Wikipedia article and recommends library books in the same NDC category to readers of the article. In the test implementation of the system, the precision of assigning NDC categories to Wikipedia articles using a convolutional neural network was as high as 87.4%, while the precision of selecting books for recommendation using a support vector machine reached 99.8%.**

Keywords—Book Recommendation; Wikipedia; University Library; Convolutional Neural Network; Support Vector Machine; Nippon Decimal Classification

I. INTRODUCTION

Wikipedia has become an important source of information for university students. It has been reported that university students tend to read and do research with Wikipedia articles than with books, even when within a library [1]. Within this context, a system was developed for recommending library books to Wikipedia readers when they are within a university library. The system is addable to the Web browser of a library desktop PC. The recommended books are displayed to the Wikipedia article reader with their covers and call numbers within that library. The system is aimed at encouraging students to read library books as a more reliable source of information rather than relying on Wikipedia articles. Although the demonstration of the proposed system in the present study considered readers of Japanese Wikipedia articles in Japanese university libraries, the system is applicable to any language.

The methodology of the proposed library system involves the following two steps:

(1) Assigning a Nippon Decimal Classification (NDC) category to each Japanese Wikipedia article.

(2) Selecting books to recommend to readers of each article.

In the context of these two steps, the present study considered two questions: (a) how can NDC categories be automatically assigned to Wikipedia articles? (b) how can books of the same NDC categories as Wikipedia articles be selected?

The reason for assigning NDC categories to Wikipedia articles is as follows. As would be further discussed, 1,070,202 Japanese Wikipedia articles were considered in this study, and the average number of printed books held by Japanese university libraries in 2017 was approximately 227,244 (= 323,595,000 books divided by 1,424 libraries reported by the Japan Library Association in 2018 [2]). If all the books in a library are considered as potential candidates for recommendation for each Wikipedia article, the total possible combination of articles and books would be 1,070,202 × 227,244. A system based on this figure would require a time-consuming and unrealistic operation. It thus becomes rational to adopt a process for reducing the number of candidate books for a particular article. The NDC is very handy for this purpose. Almost all Japanese books in Japanese libraries are assigned an NDC category as part of their respective call numbers. If NDC categories are similarly assigned to Wikipedia articles based on their content and only books of the same category are considered as recommendation candidates for a particular article, the number of possible combinations of books and articles would be significantly reduced. For instance, if a Wikipedia article is assigned NDC category 324 (i.e., Civil Law), only books in the same NDC category would be considered as recommendation candidates for that article.

A convolutional neural network (CNN) and support vector machine (SVM) were respectively used to answer the above-mentioned questions (1) and (2) in the development of the proposed system, namely, to automatically assign NDC categories to Wikipedia articles and to select books for recommendation to readers of a particular article. These two tools are representative machine learning methods for text classification [3, 4, 5, 6].

II. RELATED STUDIES

There have been some previous works on book recommendation to library users. Mikawa et al. [7] proposed the use of silhouette images captured by cameras installed in libraries and recommended books based on the gender and age of the library users. Jomsri [8] proposed the use of an association rule based on the faculty and profile of the user, their book loaning data, and book categories. The target users of the system

978-1-7281-2627-2/19/$31.00 ©2019 IEEE
DOI 10.1109/IIAI-AAI.2019.00034

are somewhat similar to those of the present system, although the present target users are further narrowed to readers of Wikipedia articles in university libraries. In addition, the information used for book recommendation in the present system mainly consists of the words in the Wikipedia article. This particularly distinguishes the system from those of Mikawa et al. [7] and Jomsri [8].

The author has previously investigated the use of books cited in Wikipedia articles and their NDC categories in making book recommendations to readers [9], and attempted to assign NDC categories to Wikipedia articles based on their setting titles, categories, and main texts using one single CNN channel [10].

III. Method

This section explains some key aspects of the proposed library book recommendations system. The first aspect is how Wikipedia articles are accessed and the distributed representation of nouns in them are determined. The second is how an NDC category is assigned to an article. The third is how books are selected for recommendation to an article reader. The fourth is obtaining the user's evaluation of the recommended books. It should be emphasized that the objective of the second aspect.

A. Accessing Wikipedia Articles and Determining the Distributed Representation of Nouns

The steps below were used to access Wikipedia articles and determine the distributed representation of nouns.

(a-1) The Japanese Wikipedia dump as of August 1, 2017 was downloaded from https://dumps.wikimedia.org/jawiki/. The file size was approximately 14 GB.

(a-2) Pages (i) ("ns" $\neq$ 0, indicating that it is not a "Main/Article") and (ii) (it is only a redirection page) of the articles were removed from the download (here, "ns" is the namespace tag). The number of pages that remained was 1,070,202, which are implied in subsequent references to Wikipedia articles.

(a-3) Mecab ver. 0.996 [11] with the mecab-ipadic-NEologd dictionary (as of August 10, 2017) [12] was used for a morphological analysis of the downloaded Wikipedia articles. A total of 5,512,620 types of nouns were identified in the articles.

(a-4) Word2Vec [13] in gensim [14] was used to obtain a 200-dimensional distributed representation of each noun in the Wikipedia articles. The parameters *window* and *min_count* in the Word2Vec module of gensim were set to 5 and 10, respectively.

B. Assigning NDC Categories to Each Wikipedia Article

This subsection explains the development of the training and testing data that were used to assign NDC categories to the Wikipedia articles. The data were inputted to the CNN and SVM.

Development of Training and Testing Data

The steps below were used to develop the training and testing data that were used to assign NDC categories to the Wikipedia articles.

(b-1) From the Wikipedia articles extracted in Step (a-2) above, the articles that contained "References" and cited books with their ISBNs were extracted. There were 50,375 such articles, including 45,151 unique ISBNs and 117,219 books (including duplicates). Only books with specified ISBNs were considered because this enabled obtaining the book bibliographies without ambiguity in Step (b-2). A procedure that utilizes book titles or author names is likely to produce differing bibliographies.

(b-2) Based on the identified book ISBNs, the unique and consistent bibliographies of the cited books were obtained using the OpenSearch tool of the National Diet Library of Japan [15], which is the largest and most comprehensive library in Japan. Bibliographies that did not contain three-digit NDCs (for instance, the three digits such as 324 indicate the book's main class 3, division 2, and section 4) or book titles were excluded. Overall, 40,647 bibliographies were obtained.

(b-3) The most popular NDC categories among the books cited in each article were identified. Regarding the level or depth of the NDC categories, the main class and division of the books were employed. For example, if an article cited five books with NDC categories 324, 324, 324, 325, and 369, respectively, the NDC category 32 (main class 3 and division 2) was considered to be the most popular, having a frequency of four. Incidentally, NDC category 32 and 36 indicate Law and Society, respectively.

(b-4) An article was extracted if the frequency of the most popular NDC category was $\geqslant$3 and its ratio relative to the total number of categories were $\geqslant$ 0.8. The threshold values of 3 and 0.8 were chosen empirically. Thus, the example article mentioned in Step (b-3) was extracted because the frequency of the NDC category 32 is 4 (> 3) and its ratio is 0.8 (= 4/5). In this way, 5,385 articles were extracted. This set of articles is henceforth referred to as the LC set.

(b-5) It was assumed that the most popular NDC category in an article in the LC set was the one that should be assigned to that article. For instance, the NDC category 32 should be assigned to the example article in Step (b-4). Accordingly, 5,385 pairs of the above-mentioned articles and NDC categories were extracted from the LC set. These were used as training and testing data for the CNN and SVM, as will be explained in the next subsection. The number of NDC categories that were to be assigned to the LC set of articles are presented in Table I, where C, N, and T respectively represent the NDC category, number of articles, and their respective total values. It can be seen from Table I that, for instance, NDC category 32 should be assigned to 107 articles in the LC set. In other words, 107 articles were obtained for the training and testing of the CNN for NDC

category 32. Although the CNN and SVM could not be assigned NDC categories that were not included in the training data, these were somewhat exceptional cases. Actually, only 16 NDC categories (05, 06,…97) were not included in the LC set.

TABLE I. NUMBERS OF NDC CATEGORIES TO BE ASSIGNED TO ARTICLES IN LC SET

C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
00	32	10	1	20	4	30	5	40	6	50	8	60	0	70	11	80	4	90	4
01	22	11	1	21	754	31	36	41	139	51	13	61	7	71	0	81	21	91	272
02	10	12	14	22	74	32	107	42	37	52	26	62	2	72	118	82	1	92	8
03	1	13	14	23	114	33	42	43	11	53	57	63	0	73	0	83	2	93	9
04	2	14	52	24	5	34	0	44	13	54	32	64	44	74	14	84	0	94	10
05	0	15	3	25	6	35	0	45	142	55	234	65	20	75	45	85	0	95	6
06	0	16	53	26	0	36	46	46	17	56	4	66	6	76	190	86	0	96	0
07	4	17	17	27	1	37	23	47	114	57	2	67	20	77	576	87	0	97	0
08	0	18	79	28	145	38	73	48	279	58	8	68	101	78	330	88	0	98	1
09	0	19	60	29	159	39	62	49	219	59	81	69	7	79	90	89	1	99	2
T	71	T	294	T	1,262	T	394	T	977	T	465	T	207	T	1,374	T	29	T	312

(b-6) The above-mentioned 5,385 pairs were randomly divided into 4,835 training data and 500 testing data. The former was used to train the CNN and SVM to assign NDC categories to each Wikipedia article, while the latter was used to compare the performances of the CNN and SVM. The better of CNN and SVM was then used to assign NDC categories to all the Wikipedia articles. As will be discussed later, the assigned NDC categories were used to select books for recommendation to the Wikipedia article readers.

Input to CNN

The input to the employed CNN utilized three channels, which correspond to the title, category, and main text of the Wikipedia articles. The structures of the channels and how the input was implemented are discussed below.

First, the nouns in the title of each Wikipedia article in the LC set were extracted and represented by their distributed representation (i.e., 200-dimensional vectors obtained through Step (a-4) in subsection III.A). The vectors and other zero vectors constitute a 50 (rows) by 200 (columns) matrix, referred to as the title channel for the CNN. In the title channel matrix, the i-th row contains (a) the 200-dimensional vector for the i-th noun in the title, when $i \leqslant 5$ (the threshold was chosen empirically), and (b) a zero vector, when $i > 5$. If the number of nouns in a certain title is less than five, the corresponding row would contain a zero vector. If the number of nouns is greater than five, only the first five nouns would be adopted in the title channel. This matrix is henceforth referred to as *tl*.

The first five categories in each LC set article were extracted and represented by a 50 × 200 matrix. The $\{(i-1)*10+j\}$-th row of the matrix contained the 200-dimensional vector (obtained by Step (a-4) in subsection III.A) for the j-th noun in the i-th category ($i \leqslant 5$ and $j \leqslant 5$). The threshold 5 was chosen empirically. The other rows contained zero vectors. If the number of nouns in a certain category was less than five, the corresponding row was set to a zero vector. If the number of nouns was greater than five, only the first five nouns in each category was adopted. This matrix is henceforth referred to as *cg*.

Finally, ten nouns within the main text of each LC set article with the highest TF-IDFs (term frequency-inverse document frequency) were extracted. The TF-IDF of noun X in article Y is defined as "frequency of X in the main text of Wikipedia article Y" multiplied by "log (number of Wikipedia articles/number of articles with main text containing X)." These were represented by a 50 × 200 matrix. The $\{(i-1)*10+1\}$-th row contained the 200-dimensional vector (obtained through Step (a-4) in subsection III.A) of the noun whose TF-IDF was the i-th highest in the main text ($i \leqslant 10$). The other rows were set to zero vectors. This matrix is henceforth referred to as *tx*.

Hence, tl, cg, and tx constituted the input channels to the present CNN. All seven possible combinations of these channels were tested, namely, (a) tl, (b) cg, (c) tx (i.e. only one channel), (d) tl and cg, (e) tl and tx, (f) cg and tx (two channels), and (g) tl, cg, and tx (all three channels).

CNN Settings

The CNN is well-known for its excellent ability for visual recognition. The above-mentioned matrices can be considered as two-dimensional (2D) images. The present CNN utilized five kinds of filters of heights 1, 2, 3, 4, and 5, respectively. All the filters had a width of 200. Furthermore, for each filter size, six different numbers of filters were used in the tests, namely, 50, 100, 150, 200, 250, and 300.

The Tensorflow ver. 1.0 [16] was used to create and train the CNN. The strides and padding of the *tf.nn.conv2d* function were set to [1, 1, 1, 1] and VALID, respectively. Rectified linear units (*tf.nn.relu*) were used after *tf.nn.conv2d*, and max pooling was employed. The ksize, strides, and padding of the *tf.nn.max_pool* function were set to [1, IH-FH+1, 1, 1], [1, 1, 1, 1], and VALID, respectively, where IH and FH respectively represent the heights of the input channel and filter. The *tf.nn.softmax* function was used in the final output layer. The size of the mini batches, number of epochs, dropout rate, and learning rate were 100, 300, 0.5, and 0.0001, respectively. The data flow is shown in Figure 1.

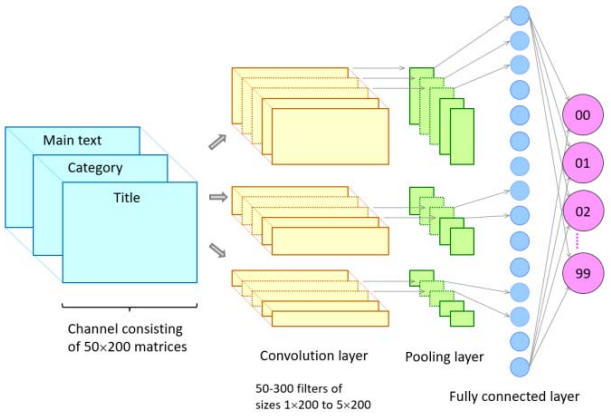

Fig. 1. Data Flow of the Developed CNN

Input to SVM

The channels mentioned in the previous subsections were converted into the vector V, where the V's {(i−1)*200+j}-th element is the (i, j) element of the channel and was inputted to the SVM. When multiple channels were used for the input to the SVM in this study, the corresponding vectors were concatenated for the input. For instance, when all the three channels (title, category, and main text) were used, the dimension of the input vector was 30,000 (= 50 × 200 × 3).

SVM Settings

The LIBSVM ver. 3.23 [17] was used with the SVM to assign NDC categories to the Wikipedia articles. The radial basis function (RBF) kernel of the SVM was employed for this purpose. The easy.py of the LIBSVM package was used to determine the optimal values of the parameters C and γ.

C. Selecting Books to Recommend

This subsection describes the selection of books to recommend to a Wikipedia article reader. The training and testing data were generated and inputted to the CNN and SVM.

Generation of Training and Testing Data

The following steps were used to generate the training and testing data for selecting books to recommend to Wikipedia article readers.

(c-1) As mentioned in section I, the aim of the present study was the development of a system for recommending library books to students who read Wikipedia articles in university libraries. The experimental implementation of the system in this study was conducted at T University. There were 621,129 candidate books for recommendation in the library, the books having been borrowed at least once between January 2, 2006 and March 31, 2017. It was assumed that books that had not been borrowed in more than 10 years would not be suitable for recommendation. The exclusion of such books also had the additional benefit of reducing the computational cost of the system.

(c-2) The distribution of the NDC categories of the 621,129 candidate books is presented in Table II, where C and N represent NDC category and number of corresponding samples, respectively. It can be seen from the Table II that, for instance, the number of books with NDC category 32 is 21,276. Incidentally, the 621,129 candidate books had 3,830,241 loan records.

TABLE II. DISTRIBUTION OF NDC CATEGORIES OF BOOKS BORROWED FROM T UNIVERSITY LIBRARY

C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
00	14,544	10	3,891	20	5,244	30	10,399	40	6,088	50	8,770	60	4,060	70	10,521	80	8,081	90	3,746
01	8,562	11	2,388	21	21,987	31	18,177	41	26,147	51	8,063	61	7,719	71	1,685	81	7,188	91	19,697
02	3,584	12	5,592	22	8,877	32	21,276	42	11,020	52	8,751	62	1,769	72	15,958	82	2,886	92	5,084
03	572	13	10,362	23	4,356	33	25,639	43	7,002	53	2,774	63	146	73	775	83	4,507	93	5,152
04	827	14	13,305	24	1,355	34	2,696	44	1,495	54	8,672	64	1,011	74	1,774	84	809	94	2,111
05	687	15	1,417	25	1,141	35	2,482	45	6,243	55	569	65	2,391	75	4,135	85	2,330	95	2,538
06	754	16	4,567	26	418	36	25,711	46	8,226	56	1,329	66	617	76	2,317	86	147	96	172
07	1,088	17	1,423	27	225	37	36,168	47	2,089	57	1,764	67	4,002	77	2,920	87	86	97	116
08	24,895	18	4,694	28	4,383	38	11,024	48	2,847	58	1,297	68	2,057	78	15,074	88	675	98	687
09	1,761	19	3,750	29	4,740	39	1,399	49	43,479	59	717	69	793	79	692	89	611	99	388
T	57,274	T	51,389	T	52,726	T	154,971	T	114,636	T	42,706	T	24,565	T	55,851	T	27,320	T	39,691

(c-3) The author, together with others, has previously reported that the recommendation of frequently borrowed books produces better results [8]. To promote the recommendation of such books, the loan frequencies of the library books were adopted as a feature in the CNN and SVM, specifically the number of times that each book had been borrowed between January 2, 2006 and March 31, 2017. This parameter is henceforth referred to as N. Each book was then assigned a value ln(N/Y+1), where Y is (a) 2018 minus the publication year of the book, if the book was published in or after 2006, or (b) 12, if the book was published before 2006. For instance, ln(20/4+1) = ln(6) was assigned to a book published in 2014 and had been borrowed 20 times during the above-mentioned period. Henceforth, ln(N/Y+1) is referred to as LF. The LF is the natural logarithm of a rough approximation of the number of times the book had been borrowed per year during the above-mentioned period. The initial idea was to use N/Y, but the above natural logarithm seemed to be more effective in the pilot experiment. Further study is planned to confirm this observation based on the responses of the subjects to questions presented to them.

(c-4) The sets {ai}, {xi}, {yi}, and {zi} were defined as distributed representations (the 200-dimensional vectors obtained in Step (a-4) in subsection III.A) of the first five nouns in the book titles (A), the first five nouns in the titles of the Wikipedia articles (X), the first five nouns in the first five categories of the Wikipedia articles (Y), and ten nouns in the main texts of the Wikipedia articles (Z) with the highest TF-IDFs, respectively. Calculations were then implemented to obtain a set of values {dot(ai, xj)}i,j, which were arranged in descending order, where dot(p, q) is the inner product of vectors p and q. If the number of values was less than a threshold M, 0's were appropriately added to the end of the sequence. This sequence of values (i.e., a vector) is henceforth referred to as SAX. The vectors SAY and SAZ were similarly obtained from {dot(ai, yj)}i,j and {dot(ai, zj)}i,j, respectively.

(c-5) Pairs of Wikipedia articles and the books they cited were extracted as training data (i.e., pairs of Wikipedia articles and books that should/should not be recommended to the article readers). More concretely, from the combination of 50,375 articles and 117,219 books they cited (mentioned in (b-1) in subsection III.B), pairs of articles and books were extracted if they satisfied all the following conditions: (1) the assigned NDC categories of the former and the NDC categories of the latter were identical, (2) LF of the books in Step (c-3) were no less than ln2 (implying that the books were borrowed at least once a year on average), and (3) the maximum values of the elements of SAX, SAY, and SAZ in Step (c-4) were $\geqslant$0.9. The threshold 0.9 was chosen heuristically through preliminary experiments. The pairs were regarded as positive examples (i.e., examples of Wikipedia articles that the corresponding books should be recommended for). A total of 5,444 such pairs of articles and books were extracted. Among these were 3,943 articles, which implies that some articles were paired with more

than one book. Articles on Leukemia and Galois Theory were the most represented and appeared in 14 pairs.

(c-6) For each of the 3,943, the books were chosen when (1) their NDC categories were identical to the NDC categories assigned to the articles and (2) they were not cited in the articles. For instance, when a certain article appeared in three positive examples, three books with the same NDC category as that assigned to the article and which were not cited in the article were randomly chosen. In this way, another 5,444 pairs of articles and books were generated and regarded as negative examples. A total of 10,888 training data was thus obtained.

Input to and Settings of CNN and SVM

The scalar LF in Step (c-3) and vectors SAX, SAY, and SAZ in Step (c-4) were concatenated and inputted to the CNN and SVM. Regarding M in Step (c-4), values of 10, 20, and 30 (inner products of vectors) were considered in the tests. Therefore, for example, for M = 30, a vector of length 91 (= 1 + (30 × 3)) was inputted to the CNN and SVM. In this case, the number of the CNN input channels was one, and the height and width of the filter were 1 and 1+M*3, respectively. The tests considered 50, 100, 200, and 400 filters, respectively. Regarding the SVM, the easy.py was once again used to determine the optimal values of C and γ for the radial basis function kernel.

In the tests, the present CNN (or SVM) was used to assign the NDC category (main class and division) of each Wikipedia article, as described in subsection III.B. Subsequently, for each Wikipedia article, all the books in the T University Library with NDCs identical with that of the article were identified. For example, for a Wikipedia article with an assigned NDC of 32, all the books in the T University Library with an NDC of 32 were identified. The three books with the highest probabilities in the CNN and SVM were then recommended for that article.

IV. Results

This section presents the results of (1) the assignment of NDC categories to Wikipedia articles, (2) the selection and recommendation of books to articles readers, and (3) the evaluation of the recommended books by articles readers.

A. Assigning NDC Categories to Wikipedia Articles

The precision of the NDC category assignment to the Wikipedia articles was defined as the ratio of the number of articles to which correct categories were assigned to the total number of articles, expressed as a percentage. The CNN-based NDC category assignment results are presented in Table III, where tl, cg, ..., tl+cg+bd respectively represent channel combinations (1)–(7) mentioned in subsection III.B. It can be seen from the table that, for example, when 100 filters and only the titles of the Wikipedia articles (i.e., tl) are used, the NDC category assignment precision is 49.0%. The use of article titles, categories, and main texts with 250 filters produces the highest precision of 87.4%. It is also obvious that the use of more filters and channels does not necessarily improve the results, owing to the problem of overfitting.

TABLE III. Precision of NDC Category Assignment Using CNN

	Number of Filters					
	50	100	150	200	250	300
tl	47.4	49.0	47.4	48.6	49.4	49.2
cg	73.4	67.6	67.8	69.2	63.4	68.0
bd	75.2	80.0	80.0	80.8	80.4	80.8
tl+cg	73.6	75.2	73.0	74.0	75.0	77.2
tl+bd	80.2	80.8	82.0	82.4	81.2	82.2
cg+bd	84.8	84.4	84.8	85.8	86.0	86.2
tl+cg+bd	85.8	85.0	86.2	86.0	87.4	85.4

Table IV presents the SVM-based NDC category assignment results for the LC set of Wikipedia articles. It can be seen from the table that, for example, the use of the distributed representation of only the titles produces a precision of just 67.2%, corresponding to C and γ (subsection III.B) values of 128 and 0.031250000, respectively. The highest precision of 84.4% is achieved using the distributed representations of the titles, categories, and main texts, with C and γ values of 512 and 0.000122070. This precision is lower than the highest value of 87.4% achieved by the CNN (Table III).

TABLE IV. Precision of NDC Category Assignment Using SVM

	Precision	C	γ
tl	67.2	128	0.031250000
cg	79.3	32	0.001953125
bd	79.1	8	0.001953125
tl+cg	80.6	128	0.001953125
tl+bd	80.1	512	0.000488281
cg+bd	84.0	512	0.000003052
tl+cg+bd	84.4	512	0.000122070

Based on the above results, the developed CNN was used to assign NDC categories to all the Wikipedia articles using their titles, categories, and main texts and 250 filters. The number of Wikipedia articles with respect to the NDC category obtained by this means is presented in Table V. Here, C and N represent the NDC category and number of articles, respectively. It can be seen that, for example, the number of Wikipedia articles assigned NDC category 32 is 9,447.

TABLE V. Number of Wikipedia Articles with Respect to Assigned NDC Category

C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
00	14,434	10	0	20	181	30	3,373	40	963	50	1,522	60	0	70	2,763	80	1,229	90	1,292
01	1,892	11	73	21	68,066	31	12,702	41	7,338	51	2,120	61	23,774	71	0	81	4,614	91	23,539
02	1,735	12	1,945	22	27,584	32	9,447	42	4,308	52	9,379	62	0	72	29,984	82	0	92	1,430
03	0	13	2,594	23	41,070	33	11,263	43	4,570	53	15,671	63	0	73	0	83	1,157	93	4,007
04	98	14	3,599	24	808	34	0	44	8,123	54	13,704	64	4,506	74	2,589	84	0	94	1,988
05	0	15	0	25	902	35	0	45	7,352	55	13,736	65	212	75	1,806	85	0	95	0
06	0	16	3,787	26	0	36	10,642	46	1,841	56	210	66	3	76	95,971	86	0	96	0
07	304	17	4,494	27	0	37	34,331	47	5,493	57	83	67	5,490	77	101,957	87	0	97	0
08	0	18	9,578	28	4,410	38	4,880	48	10,752	58	963	68	49,171	78	127,860	88	0	98	0
09	0	19	7,197	29	92,185	39	32,884	49	19,692	59	5,668	69	22,049	79	22,865	89	0	99	0
T	18,463	T	33,267	T	235,206	T	119,522	T	70,432	T	63,056	T	105,205	T	385,795	T	7,000	T	32,256

B. Selecting Books to Recommend

The precisions of the CNN and SVM for classifying books to recommend or not recommend are presented in Tables VI and VII. A comparison of the tables reveals that the SVM produces better results, with its precision reaching 99.8%. The SVM was thus used to select books for recommendation to the Wikipedia articles readers. The three books with the highest probabilities

of belonging to the group of books to be recommended for each Wikipedia article were selected and recommended to the subjects of the test implementation of the propose system.

TABLE VI. PRECISION OF BOOK CLASSIFICATION USING CNN

	Number of Filters			
M	50	100	200	400
10	98.3	98.4	98.3	98.4
20	98.4	98.7	98.6	98.5
30	98.6	98.6	98.6	98.5

TABLE VII. PRECISION OF BOOK CLASSIFICATION USING SVM

M	Precision	C	γ
10	99.8	512	0.03125
20	99.8	32	0.12500
30	99.7	32	0.12500

V. DISCUSSION

This section considers some potential error sources in the present work and how such may be addressed for further study. Among the 500 test data that were used for NDC category assignment, 63 Wikipedia articles were wrongly assigned NDC categories by the CNN. Among these 63, articles with proper nouns in their titles were predominant, with 12 containing the names of persons and six the names of organizations. Three articles with the names of historic Japanese personalities in their titles were assigned NDC category 21 (History of Japan), although the "correct" category was 28 (Biography). In this study, the correct NDC category was determined based on the books cited in the Wikipedia articles. These persons were mentioned in articles that cited their biographies. However, considering that the personalities are actually of historic significance, the above-mentioned assignment of History of Japan to the relevant articles is not entirely a failure. Other "incorrect" categorizations were observed for articles that belonged to multiple categories. For example, while the "correct" NDC category of an article on a historic Chinese calligrapher and philosopher is 72 (Painting and Calligraphy), the proposed system assigned category 12 (Oriental Philosophy) to it. This apparent failure may be difficult to correct, and it may not even be necessary to attempt a correction for the purpose of book recommendation because recommending a book for a specific aspect of the article may be acceptable.

The book selection precision of the SVM reached 99.8% and few errors were observed here.

VI. CONCLUSIONS

To encourage students to read library books as a more reliable source of information, a system was developed for recommending library books to Wikipedia article readers in university libraries. The system assigns NDC categories to Wikipedia articles and recommends library books in the same categories as the respective articles. The CNN-based NDC category assignment precision of the system for Wikipedia articles was determined to be as high as 87.4%, while the SVM-based book selection precision reached 99.8%. In the present study, the system was implemented as a Google Chrome extension. It would be interesting to further implement the system as an extension of other Web browsers such as Microsoft Edge and Mozilla Firefox, and investigate whether the books recommended to the Wikipedia article readers are actually borrowed from the library.

REFERENCES

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

[2] Japan Library Association (2018) Statistics on library in Japan. Japan Library Association. 515p.

[3] Kim, Y. (2014) Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746-1751.

[4] Arai, S. and Tsuji, K. (2015) Automatically assigning NDC categories to reference service records by using machine learning methods. *J Jpn Soc Inf Knowl* 25(1):23-40. (Text in Japanese.)

[5] Johnson, R. and Zhang, T. (2015) Effective use of word order for text categorization with convolutional neural networks. *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 103-112.

[6] Wang, P. et al. (2015) Semantic clustering and convolutional neural network for short text categorization. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 352-357.

[7] Mikawa, M. et al. (2011) Book recommendation signage system using silhouette-based gait classification. *Proceedings or the 10th International Conference on Machine Learning and Applications*, 416-419.

[8] Jomsri, P. (2018) FUCL Mining technique for book recommender system in library service. *Proceedings of the 11th International Conference Interdisciplinarity in Engineering*, 550-557.

[9] Tsuji, K. (2016) Books cited in Wikipedia: possibility to use their Nippon decimal classification categories for book recommendation. *Proceedings of the 7th International Conference on E-Service and Knowledge Management*, 1196-1197.

[10] Tsuji, K. (2017) Automatic classification of Wikipedia articles by using convolutional neural network. *Proceedings of the 9th Qualitative and Quantitative Methods in Libraries International Conference*, 8p. (No Pagination).

[11] Mecab. http://taku910.github.io/mecab/ [Last Access: 2019-01-14]

[12] mecab-ipadic-NEologd: neologism dictionary for MeCab. https://github.com/neologd/mecab-ipadic-neologd [Last Access: 2019-01-14]

[13] Word2vec. https://radimrehurek.com/gensim/models/word2vec.html [Last Access: 2019-01-14]

[14] gensim: topic modelling for humans. https://radimrehurek.com/gensim/ [Last Access: 2019-01-14]

[15] OpenSearch by the National Diet Library of Japan. http://iss.ndl.go.jp/information/api/ [Last Access: 2019-01-14]

[16] Tensorflow. https://www.tensorflow.org/ [Last Access: 2019-01-14]

[17] LIBSVM: library for support vector machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/ [Last Access: 2019-01-14]