Assigning Library Classification Numbers to People on the Web

Harumi Murakami¹, Yoshinobu Ura², and Yusuke Kataoka¹

¹ Graduate School for Creative Cities, Osaka City University,
3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 Japan
harumi@media.osaka-cu.ac.jp
http://murakami.media.osaka-cu.ac.jp/

² Winspire,
9-310, Hachiban-cho, Wakayama 640-8157 Japan

Abstract. To help users select and understand people during searches for them, we present a method of assigning Nippon Decimal Classification (NDC), which is a system of library classification numbers, to people on the web. By assigning NDC numbers to people, we can assign not only labels to people but also build a NDC-based people-search directory. We use a relative index in NDC, which lists the related index terms attached to NDC. We developed a prototype based on this approach. We evaluated the usefulness of our proposed method and directory and found that extracting relative index terms from the titles of web pages outperformed comparative methods.

Keywords: NDC, library classification system, relative index, Web people search, people-search directory.

1 Introduction

The popularity of web people searches continues to rise as the number of people increases about whom the web can provide information. Most people search systems are based on keyword search. By keyword search, which is typically a search by a person name or a keyword, users distinguish different people from the search results. If the list is merely “person 1, person 2, and so on,” users have difficulty determining which person they should select. Appropriate labels shown with people should help users select the person they want.

There is research that assigns labels to people. For example, Wan et al. separated web people search results and assigned titles to person clusters [1]. Ueda et al. assigned vocation-related information to person clusters [2]. Mori et al. extracted keywords contained in web pages [3].

In this paper, we present an approach of assigning labels to people to help users select and understand people. We use Nippon Decimal Classification (NDC), which is a library classification system in Japan, whose organization resembles the Dewey Decimal Classification (DDC). NDC is comprised of ten classes, each of which is divided into ten divisions, and each division has ten sections, and
so on. The NDC number is constructed from three digits (with other optional
digits after the decimal point.)

By assigning NDC numbers to people, we can assign labels to people and
build a NDC-based people-search directory. For example, when we assign 312.8
(Politician) to a former Japanese prime minister Naoto Kan, users can browse
300 (Social sciences: class) to 310 (Political sciences: division) to 312 (Political
history and conditions: section) and find him in the directory.

Although library classification systems were designed to classify library col-
clections instead people, we exploit their advantages because many categorization
schemes proposed for web resources lack the rigorous hierarchical structure and
careful conceptual organization found in established schemes [4] such as library
classification systems. In this paper, we use NDC, which resembles DDC both
in its organization schemes and in having relative index terms. Moreover, NDC
is the most popular library classification system in Japan. This research assigns
NDC numbers to people on the web, and develops a NDC-based people-search
directory.

Below, we explain our approach in Section 2 and examples of our implemented
prototype in Section 3. Our experiments are described in Section 4. We discuss
the significance of our research in Section 5.

2 Approach

2.1 Overview

Our approach uses a relative index in NDC. The relative index lists the related
index terms attached to NDC numbers. For example, three index terms talent, intelлект, intelligence are attached to 141.1 (Intelligence). There are 29,514 index
terms and 8,551 NDC9 (version 9) numbers.

Our proposed algorithms are constructed from two processes: (1) extracting
relative index terms from web pages, and (2) assigning NDC numbers to people
(Figure 1).

2.2 Extracting Relative Index Terms

When HTML files of a person are given, after removing the HTML tags, we ex-
tract the relative index terms from the texts inside the title tags. When multiple
index terms can be extracted, the longest-match method is used.

We deleted the following index terms that we consider unnecessary: (a) those
that consist of one character, and (b) 100 manually selected terms that often
appear on the web.

2.3 Assigning NDC Numbers

After the index terms are converted to NDC numbers, they are assigned based
on the following scores:
Fig. 1. Overview of the algorithm

\[
\text{score}(\text{ndc}_i) = \frac{\text{freq}(\text{ndc}_i)}{\sum_{k=1}^{n} \text{freq}(\text{ndc}_k)}
\]  

Where ndc is a NDC number and \( n \) is a distinct number of NDC numbers attached to a person.

2.4 Example

Consider the following sentence: “Suguru Egawa is a former professional baseball player and a baseball commentator. Metal bats have been introduced in high-school baseball...” Former, professional baseball, baseball, metal, and baseball are extracted as index terms. Former is removed because it consists of just one Japanese character. Professional baseball and baseball are converted to 783.7 (Baseball), and metal is converted to 524.28 (Metal. Alloy. Architectural hardware) and 560 (Metal engineering. Mine engineering).

The scores of 783.7 are 0.6 (3/5), 524.28 and 560 are 0.2 (1/5), respectively. These numbers can be attached to Suguru Egawa. For the top ranked score 783.7 (Baseball), its class is 700 (The arts. Fine arts), its division is 780 (Sports and physical training), and its section is 783 (Ball games).

3 Prototype

We implemented a prototype using our proposed method. This is an example of assigning the top five NDC numbers using the title documents in a dataset (see Section 4).
Figure 2 shows an initial screen of a NDC-based people-search directory. When a user selects 780 (Sports and physical training), Figure 3 is displayed. The upper side of the screen lists the list of the divisions of 780, and the lower side of screen shows the list of people assigned to 780. For example, Suguru Egawa (former baseball player) and Ai Fukuhara (table tennis player) are displayed, with five NDC numbers assigned to each. When a user selects a person, information about him or her (in this case, the search result pages of the designated person) is displayed.

<table>
<thead>
<tr>
<th>Class</th>
<th>Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 General works</td>
<td>000 General works</td>
</tr>
<tr>
<td>010 Libraries, library and inform.</td>
<td>010 Libraries, library and inform.</td>
</tr>
<tr>
<td>020 Works bibliography</td>
<td>020 Works bibliography</td>
</tr>
<tr>
<td>030 General encyclopedia</td>
<td>030 General encyclopedia</td>
</tr>
<tr>
<td>040 General collected essays</td>
<td>040 General collected essays</td>
</tr>
<tr>
<td>100 Philosophy</td>
<td>100 Philosophy</td>
</tr>
<tr>
<td>110 Special topics on etc.</td>
<td>110 Special topics on etc.</td>
</tr>
<tr>
<td>120 Oriental thought</td>
<td>120 Oriental thought</td>
</tr>
<tr>
<td>130 Western philosophy</td>
<td>130 Western philosophy</td>
</tr>
<tr>
<td>140 Psychology</td>
<td>140 Psychology</td>
</tr>
<tr>
<td>200 General history</td>
<td>200 General history</td>
</tr>
<tr>
<td>210 General history of Japan</td>
<td>210 General history of Japan</td>
</tr>
<tr>
<td>220 General history of Asia</td>
<td>220 General history of Asia</td>
</tr>
<tr>
<td>230 General history of Europe</td>
<td>230 General history of Europe</td>
</tr>
<tr>
<td>240 General history of Africa</td>
<td>240 General history of Africa</td>
</tr>
<tr>
<td>300 Social sciences</td>
<td>300 Social sciences</td>
</tr>
</tbody>
</table>

Fig. 2. Initial screen

<table>
<thead>
<tr>
<th>Division</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 General history</td>
<td>200 General history</td>
</tr>
<tr>
<td>210 General history of Japan</td>
<td>210 General history of Japan</td>
</tr>
<tr>
<td>220 General history of Asia</td>
<td>220 General history of Asia</td>
</tr>
<tr>
<td>230 General history of Europe</td>
<td>230 General history of Europe</td>
</tr>
<tr>
<td>240 General history of Africa</td>
<td>240 General history of Africa</td>
</tr>
</tbody>
</table>

Search results for "Sports and physical training"

<table>
<thead>
<tr>
<th>Name</th>
<th>Class</th>
<th>Division</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suguru Egawa</td>
<td>780 Sports and physical training</td>
<td>781 Gymnastics, etc.</td>
<td>781 Gymnastics, etc.</td>
</tr>
<tr>
<td>Ai Fukuhara</td>
<td>780 Sports and physical training</td>
<td>782 Track and field athletics</td>
<td>782 Track and field athletics</td>
</tr>
</tbody>
</table>

Fig. 3. Screen list of people

Bold letters indicate category (division, in this case)
4 Experiment

4.1 Dataset

We describe a previously developed dataset [5]. The twenty person names used in related work [6] were selected as queries. 100 web pages (HTML files) were obtained for each twenty queries from web searches (i.e., \(20 \times 100 = 2,000\) HTML files). We manually classified these web pages into different people. 152 people were found in all 2,000 web pages.

4.2 Experiment 1

We evaluated the usefulness of our algorithm that assigns NDC numbers to people (person clusters).

Method. We assigned NDC numbers to people (person clusters) with three methods using the following six documents (i.e., \(3 \times 6 = 18\)): (a) Tf-idf, (b) Cosine, and (c) Our method. The six documents were (1) Title, (2) Html, (3) Snippet, (4) Kwic50, (5) Kwic100, and (6) Kwic200.

The Tf-idf and Cosine methods do not use relative index terms. We treated a document of a person cluster as a query and a NDC label as a document. The numerator for calculating idf is the total amount of NDC numbers.

The Title is a document extracted from the title elements. The Html is entire document. The Snippet is a document given as a result of a Yahoo! search. In this paper, to examine co-occurrence information, we introduce concatenated text strings before and after person names. We call this “keyword in context (kwic)”. Kwic50 is a concatenated document of 50 Japanese characters before a person’s name and 50 Japanese characters after it (i.e., \(50 + 50 = 100\) Japanese characters). Kwic100 and Kwic200 are constructed in the same way except for using \(100 + 100 = 200\) or \(200 + 200 = 400\) Japanese characters. We removed the HTML tags from all documents. Figure 4 shows an example of the six documents.

We manually selected the most appropriate NDC numbers for each person (137 people out of 152). When there is no appropriate NDC number, we set it to “none.” (i.e., \(152 - 137 = 15\) people.)

We checked whether the correct and assigned numbers (top ranked score) are the same in each class level (0-9), each division level (00-99), each section level (000-999), and the exact number. For example, when the correct NDC number is 783.7, the assigned number that starts with 7 (e.g. 700) is judged correct in the class level, which starts with 78 (e.g. 780) is judged correct in the division level, which starts with 783 (e.g. 783) is judged correct in the section level, and only 783.7 is judged correct at the exact level.

The following are the evaluation measures:

\[
\text{Precision} = \frac{\text{correct answers by the method}}{\text{people to whom NDC was assigned by the method}} \quad (2)
\]

\[
\text{Recall} = \frac{\text{correct answers by the method}}{\text{people to whom NDC was assigned manually}} \quad (3)
\]
The commissioner announced that a former Giant, Suguru Egawa (54), who is currently a baseball commentator, has been chosen as the next manager of Lotte... (example) Kwic10 (10 + 10 = 20 Japanese characters)

Baseball News
23 Jan 2011 ... Baseball News from the net. PR. Pagetop.
Baseball Suguru Egawa said about Yuki Saito: "He will win around 200 games." Yuki Saito (Waseda U), Nippon Ham’s top draft pick, ...
.../blog-entry-1376.html

Fig. 4. Six documents for evaluation

\[
F - measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)
\]

\[
\text{Accuracy} = \frac{\text{correct answers}}{\text{people}} \quad (5)
\]

When calculating the Accuracy, \textit{none} is judged correct when there is no correct NDC number for the people.

Results and Analysis. Table 1 shows the average Accuracy values in Experiment 1. Our method outperformed the comparative methods, suggesting its usefulness with relative index. For Accuracy, except for the class level, the Title was best among all the documents.

Table 2 shows the results for the exact numbers. Our methods were better than the comparative methods. In the six documents in our method, Title had good Precision and Accuracy, Snippet and Kwics had good Recall, and Kwic50 slightly outperformed the Title in the F-measure.

4.3 Experiment 2

Experiment 1 showed the overall effectiveness of our method using title documents. We evaluated the precision of our algorithm from another perspective using five-scale values.

Method. We evaluated whether the assigned NDC numbers (top ranked scores) were related to people by checking web pages by five values (5: very related; 4: slightly related 3: neutral; 2: not very related ; 1: unrelated).
Table 1. Result of Experiment 1: Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Document</th>
<th>Class</th>
<th>Division</th>
<th>Section</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tf-idf</td>
<td>Max</td>
<td>0.44 (Title)</td>
<td>0.34 (Title)</td>
<td>0.23 (Title)</td>
<td>0.15 (Title)</td>
</tr>
<tr>
<td>Cosine</td>
<td>Max</td>
<td>0.43 (Title, Kwic200)</td>
<td>0.34 (Title)</td>
<td>0.25 (Title)</td>
<td>0.17 (Title)</td>
</tr>
<tr>
<td>Our method</td>
<td>Title</td>
<td>0.52</td>
<td>0.45</td>
<td>0.36</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Html</td>
<td>0.45</td>
<td>0.36</td>
<td>0.27</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Snippet</td>
<td>0.45</td>
<td>0.36</td>
<td>0.27</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Kwic50</td>
<td>0.52</td>
<td>0.42</td>
<td>0.29</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Kwic100</td>
<td>0.51</td>
<td>0.37</td>
<td>0.28</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Kwic200</td>
<td>0.47</td>
<td>0.37</td>
<td>0.26</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Note: for Tf-idf and Cosine, the maximum values were described. They all used titles.

Table 2. Result of Experiment 1: Exact level

<table>
<thead>
<tr>
<th>Method</th>
<th>Document</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tf-idf</td>
<td>Max</td>
<td>0.08 (Title)</td>
<td>0.08 (Snippet)</td>
<td>0.07 (Snippet)</td>
<td>0.15 (Title)</td>
</tr>
<tr>
<td>Cosine</td>
<td>Max</td>
<td>0.12 (Kwic200)</td>
<td>0.10 (Html)</td>
<td>0.12 (Kwic200)</td>
<td>0.17 (Title)</td>
</tr>
<tr>
<td>Our method</td>
<td>Title</td>
<td>0.18</td>
<td>0.12</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Html</td>
<td>0.11</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Snippet</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Kwic50</td>
<td>0.16</td>
<td>0.14</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Kwic100</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Kwic200</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Note: for Tf-idf and Cosine, the maximum values were described.

Table 3. Result of Experiment 2: Relatedness

<table>
<thead>
<tr>
<th>Title</th>
<th>Html</th>
<th>Snippet</th>
<th>Kwic50</th>
<th>Kwic100</th>
<th>Kwic200</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.41</td>
<td>2.87</td>
<td>2.77</td>
<td>3.15</td>
<td>3.02</td>
<td>2.91</td>
</tr>
</tbody>
</table>

Results and Analysis. Table 3 shows the average values of relatedness. Title was best again (3.41).

From the above results of Experiments 1 and 2, we consider Title was best among six documents to extract relative index terms to assign NDC numbers to people.

4.4 Experiment 3

We investigated how many NDC numbers should be assigned to people using Title to develop a NDC-based people-search directory.

Method. We evaluated whether the top ten assigned NDC numbers were related to people by checking web pages by five values (5: very related; 4: slightly related; 3: neutral; 2: not very related; 1: unrelated).
Results and Analysis. Table 4 shows the cumulative relatedness for each rank in Experiment 3. For example, the average value of top ranked numbers was 3.43, and the top and second ranked numbers was 3.23. The average value of the top five ranked numbers exceeded 3.

We use NDC numbers not only to categorize people in a directory but also to display labels for them. The values of the top one or two are obviously better than the top three to five; however, if we only choose the top one or two, too little information is provided by the labels. We analyzed the top three to five ranked numbers and believe they are appropriate.

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.43</td>
<td>3.23</td>
<td>3.17</td>
<td>3.12</td>
<td>3.04</td>
<td>3.00</td>
<td>2.98</td>
<td>2.94</td>
<td>2.91</td>
<td>2.89</td>
</tr>
</tbody>
</table>

In this paper, we extracted the top five ranked numbers to build a directory.

4.5 Experiment 4

We investigated how many people were found and the correct rates in each category (division) to evaluate the category’s potential.

Method. We counted how many people were found in each 100 category (division): 000-990. We also counted the correct people in each 100 category to calculate the correct rate.

Results and Analysis. Figure 5 shows the Experiment 4 result. Html was the best for the number of people (5.47 people) and Title had the highest correct rate (39%).

![Fig. 5. Number of people and correct rate by division](image-url)
4.6 Experiment 5

We investigated the usefulness of our developed prototype using 14 subjects.

**Method.** Our subjects were 14 undergraduate and postgraduate males whose average age was 22.8.

Since comparing six directories is complicated for the subjects, we chose three documents for our experiment: Title, Html, and Kwic50. We did not choose Snippet because it showed no advantage from previous experiments, and Kwic50 was chosen from Kwics, because it seemed the best.

We asked them three questions: Q1, Q2, and Q3.

(Q1) Is the NDC number attached to the person appropriate? (3: appropriate; 2: partially appropriate; 1: inappropriate).

The subjects evaluated pairs of NDC numbers and a person included in each ten category (division): 000 - 900 by checking HTML files. We used 110, 310, and 810 for 100, 300, and 800 because there was no people in categories 100, 300, and 800. We calculated the averages for each category.

(Q2) Is the list of people appropriate for each category? (3: appropriate; 2: partially appropriate; 1: inappropriate).

After question 1, the subjects evaluated the same ten lists in ten categories (000 - 900). We calculated the averages by each category.

(Q3) Rank the three people directory methods and explain why.

Finally, we asked the subjects for their overall comments.

**Results and Analysis.** Figure 6 shows the question results.

For the Title, the average values for Q1 were 2.01, 2.07 for Q2 and 1.50 for Q3. For all the questions, our prototype developed using Title was the best.

71% (10/14) of the subjects ranked Title best for Q3. The following are comments from two subjects who ranked Title best in Q3: “There is little useless information and its system is easy to understand.” “The classification precision is good.”
5 Related Work and Discussion

5.1 Related Work

The initial idea of assigning library classification numbers to people was presented in a very short position paper [7]. In this new paper, we explain the algorithm and prototype in details and evaluate our algorithm and our developed prototype. In addition, we discuss the similarities and differences from various related work.

There is research that assigns labels to people. Wan et al. assigned titles (including vocations) [1], Ueda et al. assigned vocation-related information [2], Mori et al. assigned keywords to person clusters [3], and [5] extracts location information to people. WePS-2/3 conducted competitive evaluation on person attribute extraction on web pages [8]. No such research has assigned library classification numbers to person clusters.

Some research suggests NDC numbers or other terms in libraries. Kiyota et al. suggests LCSH subject headings and NDC numbers [9], and Ueda et al. suggests BSH subject headings and NDC numbers according to user input. They use web information sources as Wikipedia for matching without using relative index terms.

The automated subject classification of web documents is not new. Golub reviewed approaches to automated subject classification of textual web documents in different research communities (machine learning, information retrieval, and library science) and classified them into four categories: text categorization, document clustering, document classification, and mixed approach [10]. Our work belongs in the document classification category because we employed well-developed controlled vocabularies. Document classification is a library science approach.

Jenkins et al. organized web resources by DDC using simple classifiers [11]. They used a DDC thesaurus to match terms in documents. OCLC Scorpion is a well-known project that assigns DDC to web resources [12]. Our work resembles their approach because it compares the selected terms from documents to be classified with the terms in the vocabulary.

Frank et al. predicted Library of Congress Classifications (LCC) from LCSH subject headings and built an LCC browsing interface for a database of scholarly Internet resources [13]. They present a machine learning technique to assign LCC numbers to LCSH subject headings. This work is classified into the forth category, a mixed approach [10]. They did not evaluate their interface.

5.2 Discussion

Our experimental results show that, among six documents, Title had the best performance assigning NDC numbers to people on the web and developing a web people-search directory.

We believe that our work’s main contribution is its successful assignment of library classification numbers to people on the web for displaying labels and
building a people-search directory. To the best of our knowledge, this is the first research that assigns library classification numbers to people on the web.

Our paper also presents the titles of web pages as good sources to form virtual documents that represent people, which it does better than whole pages, kwic documents, or snippets. The kwic concept resembles window size. In expert searches, window sizes capture the proximity of terms and candidate mentions in documents [4]. Our finding is quite different from expert searches. This reflects the difference between the two tasks and provides new insights for web people searches and other types of people searches.

Although our research is limited to NDC and Japanese, our approach is easily applicable to other classification systems, such as DDC with similar organization and relative indexes or other terminology. People are one representative entity, and our approach can be applied to such entities as industries or place names.

6 Conclusions

To help users select and understand people, our method assigns Nippon Decimal Classification (NDC) to people on the web. We developed a prototype based on this approach and evaluated the usefulness of our proposed method and directory. Extracting relative index terms from the titles of web pages outperformed comparative methods.

Future work includes improving our algorithms for assigning NDC numbers to people. Second, we need to develop other kinds of datasets (e.g., more people, or famous/not famous people, etc.) to examine the effectiveness of the proposed method.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 22500219, 25330385.

References

6. Sato, S., Kazama, K., Fukuda, K.: Distinguishing between People on the Web with the Same First and Last Name by Real-world Oriented Web Mining. IPSJ Transactions on Databases 46(8), 26–36 (2005)


