Assigning Library Classification Numbers to People on the Web

Harumi Murakami, Yoshinobu Ura, Yusuke Kataoka

Citation	AIRS 2013: Information Retrieval Technology, pp.464-475
Symposium	9th Asia Information Retrieval Societies Conference, AIRS 2013, Singapore, December 9-11, 2013. Proceedings
Part of book series	Lecture Notes in Computer Science (LNCS, volume 8281)
Issue Date	2013
Type	Conference paper
Textversion	author
Relation	This is a post-peer-review, pre-copyedit version of a Conference paper published in "AIRS 2013: Information Retrieval Technology" pp.464-475. The final authenticated version is available online at: https://doi.org/10.1007/978-3-642-45068-6_40.
DOI	10.1007/978-3-642-45068-6_40

Murakami H., Ura Y., Kataoka Y. (2013) Assigning Library Classification Numbers to People on the Web. In: Banchs R.E., Silvestri F., Liu TY., Zhang M., Gao S., Lang J. (eds) *Information Retrieval Technology. AIRS 2013. Lecture Notes in Computer Science*, vol 8281. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-45068-6_40

Assigning Library Classification Numbers to People on the Web

Harumi Murakami[1], Yoshinobu Ura[2], and Yusuke Kataoka[1]

[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.

Key words: 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 collections 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*, *intellect*, *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 extract 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:

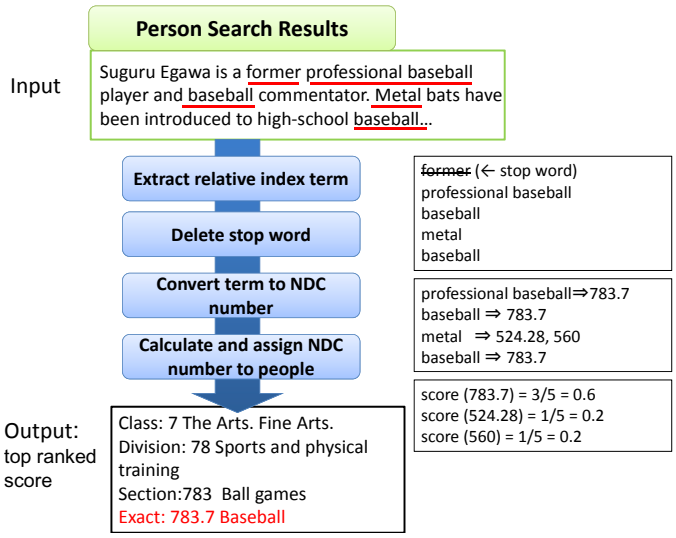

Fig. 1. Overview of the algorithm

$$score(ndc_i) = \frac{freq(ndc_i)}{\sum_{k=1}^{n} freq(ndc_k)} \tag{1}$$

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.

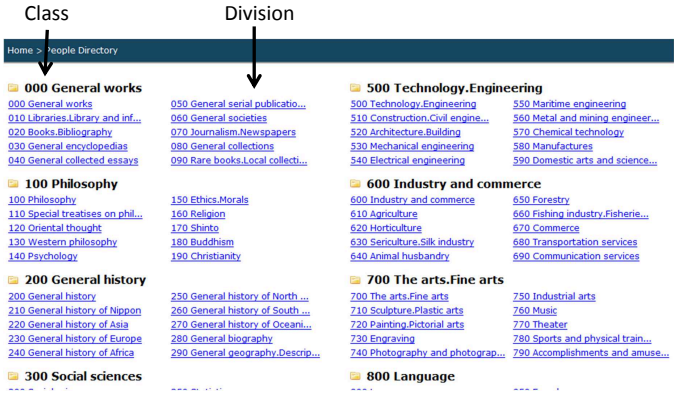

Fig. 2. Initial screen

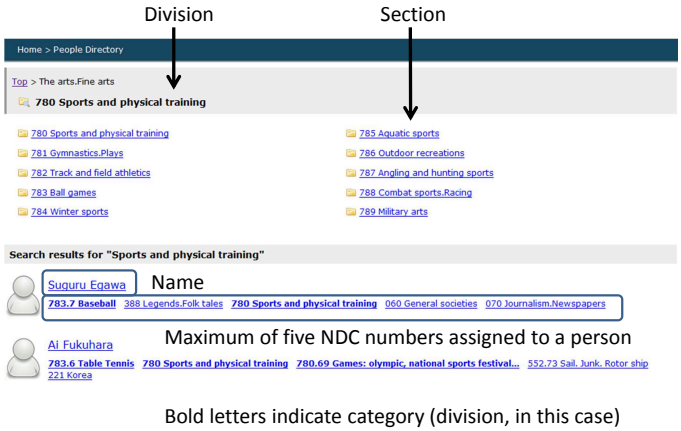

Fig. 3. Screen list of people

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). Kwic 100 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:

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

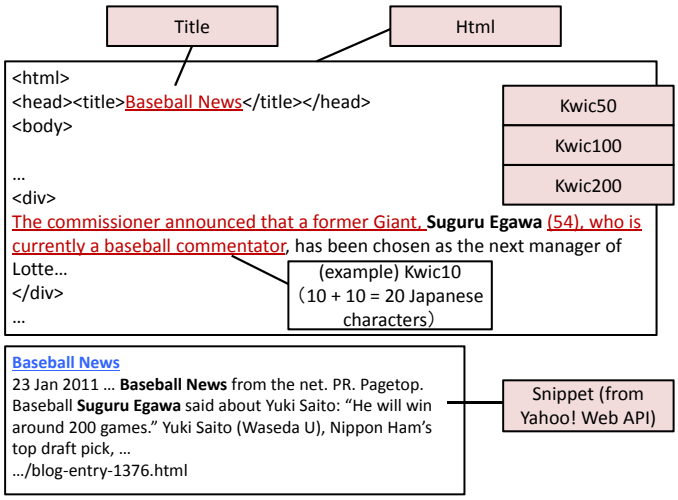

Fig. 4. Six documents for evaluation

$$Recall = \frac{\text{correct answers by the method}}{\text{people to whom NDC was assigned manually}} \tag{3}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{\text{correct answers}}{\text{people}} \tag{5}$$

When calculating the Accuracy, *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.

Table 1. Result of Experiment 1: Accuracy

Method	Document	Class	Division	Section	Exact
Tf-idf	Max	0.44 (Title)	0.34 (Title)	0.23 (Title)	0.15 (Title)
Cosine	Max	0.43 (Title, Kwic200)	0.34 (Title)	0.25 (Title)	0.17 (Title)
Our method	Title	0.51	**0.45**	**0.36**	**0.25**
	Html	**0.52**	0.43	0.29	0.12
	Snippet	0.45	0.36	0.27	0.20
	Kwic50	**0.52**	0.42	0.29	0.23
	Kwic100	0.51	0.37	0.28	0.20
	Kwic200	0.47	0.37	0.26	0.16

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

Table 2. Result of Experiment 1: Exact level

Method	Document	Precision	Recall	F-measure	Accuracy
Tf-idf	Max	0.08 (Title)	0.08 (Snippet)	0.07 (Snippet)	0.15 (Title)
Cosine	Max	0.12 (Kwic200)	0.10 (Html)	0.12 (Kwic200)	0.17 (Title)
Our method	Title	**0.18**	0.12	0.15	**0.25**
	Html	0.11	0.13	0.12	0.12
	Snippet	0.15	**0.14**	0.14	0.20
	Kwic50	0.16	**0.14**	**0.15**	0.23
	Kwic100	0.15	**0.14**	0.14	0.20
	Kwic200	0.13	**0.14**	0.14	0.16

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

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).

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.

Table 3. Result of Experiment 2: Relatedness

Title	Html	Snippet	Kwic50	Kwic100	Kwic200
3.41	2.87	2.77	3.15	3.02	2.91

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 4. Cumulative relatedness for each rank

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
3.43	3.23	3.17	3.12	3.04	3.00	2.98	2.94	2.91	2.89

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%).

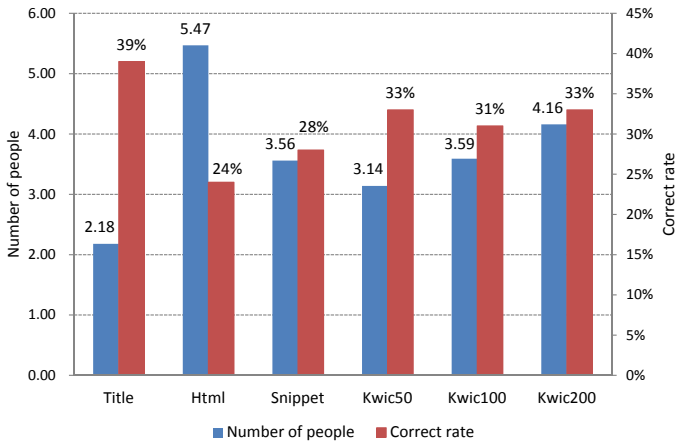

Fig. 5. Number of people and correct rate by division

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.

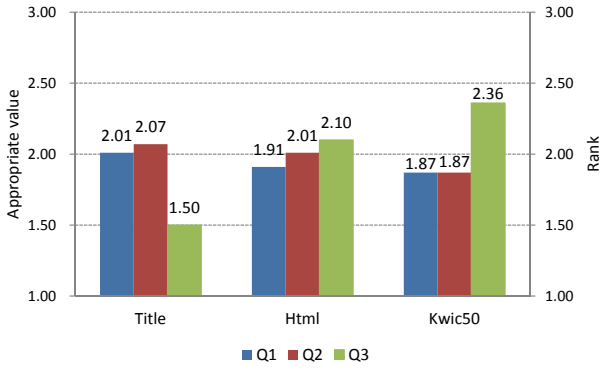

Fig. 6. Category evaluation by 14 subjects

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 [14]. 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

1. Wan, X., Gao, J., Li, M., Ding, B.: Person Resolution in Person Search Results: WebHawk. In: Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005), pp. 163–170. ACM Press, New York (2005)
2. Ueda, H., Murakami, H., Tatsumi, S.: Assigning Vocation-Related Information to Person Clusters for Web People Search Results. In: Proceedings of the 2009 Global Congress on Intelligent Systems (GCIS 2009), vol. 4, pp. 248–253. IEEE Press, New York (2009)
3. Mori, J., Matsuo, Y., Ishizuka, M.: Personal Keyword Extraction from the Web. Journal of Japanese Society for Artificial Intelligence. 20, 337–345 (2005)
4. Chan, L. M.: Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources: Issues and Challenges. In: Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium, pp. 159–178. Library of Congress, Washington DC (2001)
5. Murakami, H., Takamori, Y., Ueda, H., Tatsumi, S.: Assigning Location Information to Display Individuals on a Map for Web People Search Results. In: Song, D., Lin, C.-Y., Aizawa, A., Kuriyama, K., Yoshioka, M., Sakai, T. (Eds.) AIRS 2009. LNCS, vol. 5839, pp. 26–37. Springer, Heidelberg (2009)
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)
7. Murakami, H., Ura, Y.: People Search using NDC Classification System, In: Proceedings of Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2011), pp. 13–14. ACM Press, New York (2011)
8. Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., Amigo, E.: WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In: CLEF 2010 (2010)
9. Kiyota, Y., Nakagawa, H., Sakai, S., Mori, T., Masuda, H.: Exploitation of the Wikipedia Category System for Enhancing the Value of LCSH. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 411, ACM Press, New York (2009)
10. Golub, K.: Automated Subject Classification of Textual Web Document. Journal of Documentation. 62(3), 350–371 (2006)
11. Jenkins, C., Jackson, M., Burden, P, Wallis, J.: Automatic Classification of Web Resources using Java and Dewey Decimal Classification. Computer Networks and ISDN Systems. 30, 1-7, pp. 646–648 (1998)
12. Automatic Classification Research at OCLC, `http://www.oclc.org/research/activities/auto_class.html`
13. Frank, E., Paynter, G. W.: Predicting Library of Congress Classifications From Library of Congress Subject Headings. Journal of the American Society for Information Science and Technology. 55(3), 214–227 (2004)
14. Balog, K. Fang, Y., de Rijke, M., Serdyukov, P., Si, L.: Expertise Retrieval. Foundations and Trends in Information Retrieval. 6, 2-3, 160 (2012)