People Search using NDC Classification System

Harumi Murakami
Osaka City University
3-3-138, Sugimoto, Sumiyoshi,
Osaka 558-8585 Japan
+81-6-6605-3375
harumi@media.osaka-cu.ac.jp

Yoshinobu Ura
winspire
9-313, Hachiban-cho,
Wakayama 640-8157 Japan
+81-73-425-1215
ura@winspire.jp

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 directory. We developed a prototype based on this approach.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models.*

General Terms
Design, Algorithms

Keywords
People search directory, NDC, relative index, Web people search

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 distinguish 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 another optional digits after the decimal point.)

ESAIR'11, October 28, 2011, Glasgow, Scotland, UK.
ACM 978-1-4503-0958-5/11/10.

By assigning NDC numbers to people, we can assign labels to people and build a NDC-based person directory. For example, when we assign *312.8* (Politician) to Japanese prime minister *Naoto Kan*, users can browse *300* (Social sciences: class) to *310* (Political sciences: section) to *312* (Political history and conditions: division) and find him in the NDC-based directory. Although the NDC system was designed to classify library collections (mainly books) and not people, we exploit its advantages for the following reasons: (1) There are few established general classification system other than library classification systems in Japan, and (2) It 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.

2. APPROACH
2.1 Overview
The main feature of our approach is using 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 of Web pages, and (2) assigning NDC numbers to people.

2.2 Extracting Index Terms
When HTML files of a person are given, after removing the HTML tags, we extract the 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
Index terms are converted to NDC Numbers, which are assigned based on the following scores:

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

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. He likes wine..." *Former*, *professional baseball*, *baseball*, and *wine* are

extracted as index terms. *Former* is removed because it consists of just one Japanese character. *Professional baseball*, *baseball*, and *wine* are converted to *783.7*, *783.7*, and *588.58*. The scores of *783.7* (baseball) and *588.58* (brewage) are 0.666 (2/3) and 0.333 (1/3), respectively. These numbers are attached to "Suguru Egawa." For *783.7* (baseball), its class is *700* (The arts. Fine arts), its section is *780* (Sports and physical training), its division is *783* (Ball games).

3. PROTOTYPE

We implemented a prototype using our proposed method. We assigned five NDC9 numbers to the 137 people (person clusters) obtained in [4].

Figure 1 shows an initial screen of a NDC-based people-search directory. When a user selects *780* (Sports and physical training), Figure 2 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.

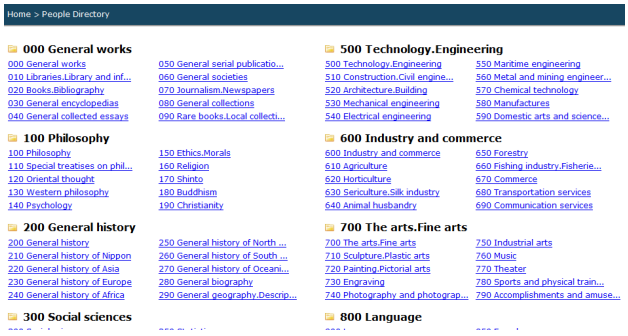

Figure 1. Initial screen.

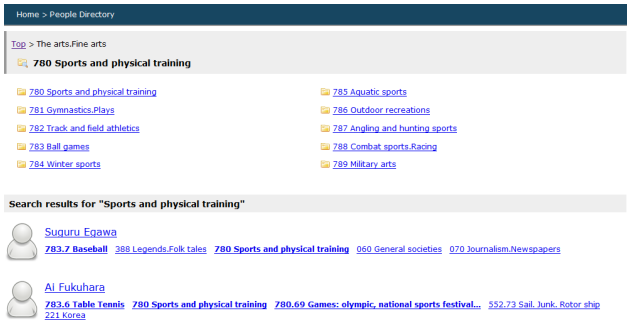

Figure 2. Screen list of people.

4. RELATED WORK AND DISCUSSION

There is research that assigns labels to people. Wan et al. assigned titles [1], Ueda et al. assigned vocation-related information [2], and Mori et al. assigned keywords to person clusters [3]. WePS-2/3 conducted competitive evaluation on person attribute extraction on Web pages [5]. 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 [6], and Ueda et al. suggests BSH subject headings and NDC numbers according to user input. They all use web information sources as Wikipedia without using relative index terms.

We believe that our main contribution is to assign library classification numbers to people to display labels and build a people-search directory. Our research is an initial step. Its proposed algorithms are simple, and we will improve them in the future. Although our research is limited to NDC and Japanese, our approach is easily applicable to other classification systems such as DDC and UDC with similar organization and relative indexes or other terminology. People are one representative *entity*, and our approach can be applied to other kinds of entities, such as industries or place names.

Although the idea of using a library classification system itself is not so new, we believe that the time has come to re-evaluate its usefulness as semantic annotations.

5. SUMMARY

We presented a method that assigned NDC numbers to people on the Web. We developed a NDC-based people search directory.

6. ACKNOWLEDMENTS

This work was supported by KAKENHI (22500219).

7. REFERENCES

[1] Wan, X., Gao, J., Li, M., and Ding, B. 2005. Person Resolution in Person Search Results: WebHawk, Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005), 163-170.

[2] Ueda, H., Murakami, H., and Tatsumi, S. 2009. Assigning Vocation-Related Information to Person Clusters for Web People Search Results. Proceedings of the 2009 Global Congress on Intelligent Systems (GCIS 2009), 4, 248-253.

[3] Mori, J., Matsuo, Y., and Ishizuka, M. 2005. Personal Keyword Extraction from the Web. Journal of Japanese Society for Artificial Intelligence. 20, 337-345.

[4] Murakami, H., Takamori, Y., Ueda, H., and Tatsumi, S. 2009. Assigning Location Information to Display Individuals on a Map for Web People Search Results, Proceedings of The Fifth Asia Information Retrieval Symposium (AIRS 2009), LNCS 5839, 26-37.

[5] Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., and Amigó, E. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. CLEF 2010.

[6] Kiyota, Y., Nakagawa, H., Sakai, S., Mori, T., and Masuda, H. 2009. Exploitation of the Wikipedia Category System for Enhancing the Value of LCSH (JCDL 2009), 411.