A System for Exploring People on the Web Using Subject Headings

Yoshitaka Shirakawa
Graduate School of Informatics
Osaka Metropolitan University
Osaka, Japan
xuanweiyue@gmail.com

Masayuki Shimokura
Graduate School of Informatics
Osaka Metropolitan University
Osaka, Japan
st23953q@st.omu.ac.jp

Takashi Hirata
Faculty of Liberal Arts
The Open University of Japan
Tokyo, Japan
takash-h@sea.plala.or.jp

Harumi Murakami
Graduate School of Informatics
Osaka Metropolitan University
Osaka, Japan
harumi@omu.ac.jp

***Abstract*—We investigated methods of assigning National Diet Library Subject Headings (NDLSH) to people for searching and identifying them on the web. We used pattern-matching and combined the following: (a) the top ten pages, (b) 100 characters before and after a person's name, (c) no synonyms, and (d) the document frequency and assigned ten NDLSHs to each person. We developed a prototype system that visually explores NDLSHs and people through network representation.**

***Index Terms*—NDLSH, web people search, exploratory search, visualization by network**

I. Introduction

Seeking information about individuals is a very popular task in web searches. Unfortunately, finding/identifying a person on the web is difficult, especially when digging through similar or identical names. When searching for information with which to identify a person, he/she must be characterized by a label. In this research, NDLSH, the National Diet Library Subject Headings (https://id.ndl.go.jp/information/download/) are assigned to people as labels. NDLSH's subject headings are the most authoritative in Japan. By assigning NDLSHs to a person on the web, well-formed keywords are assigned and exploratory searches can be conducted using broader, narrower, and related terms.

Shimokura and Murakami [1] previously investigated methods that assined NDLSH to people on the web and identified one method with the best top-ranked results. However, to find and identify people, multiple subject headings must be assigned to them. In this paper, we present methods that assign multiple NDLSHs to people on the web and develop a prototype system that visually explores NDLSH-connected people through network representation.

Below we describe our method that assigns multiple NDLSHs to people in Section 2 and our prototype system in Section 3.

II. Approach

A. Previous work

Shimokura and Murakami [1] investigated methods for assigning NDLSHs to people on the web and evaluated them using the highest results. Our study builds on this previous work, which we discuss next.

1) Dataset: They used 20 Japanese names as queries from a related work [2] and obtained 50 web pages per query from the Google Custom Search API. They manually classified these pages into different people and identified 80 separate individuals. NDLSHs were assigned to HTML files for each person.

2) Methods: First, they extracted headings with variant terms (hereafter synonyms) and deleted the following: those with two or fewer single-byte alphanumeric characters, those with only one double-byte character, those containing (...), and those containing "−−(two hyphens)" because they are less important terms and/or relatively useless for text-matching.

Since longer strings provide more specific meaning, this previous work counted the terms (headings and synonyms) from among the longer ones in the HTML documents without tags under the following conditions: (a) and (b). For example, when processing the "artificial intelligence" character string in a document, the term "artificial intelligence" is counted but not the word "intelligence." After counting the terms (headings and synonyms), the scores of the headings are calculated.

They prepared the following four types of combination conditions:

(a) Search rankings of web pages: five different patterns of the top 1, 3, 5, 10, and all the pages for each person.
(b) Positions in HTML documents: nine patterns of title only, full text, and the 20, 40, 60, 80, 100, 150, or 200 characters before and after a person's name.
(c) Synonyms: three patterns without synonyms, using synonyms with identical weights as the headings, and using synonyms with a half of the headings.

(d) Document frequency of the headings and synonyms: three patterns, doing nothing, multiplying document frequency (df) / total number of used documents (N), and multiplying the total number of used documents (N) / document frequency (df).

When these conditions are combined, they become $5 \times 9 \times 3 \times 3 = 405$ patterns.

A heading with the highest score is assigned to the corresponding person. If no heading is assigned, the answer becomes "none".

3) Evaluation of top results: They manually selected the most appropriate headings for each person (79 out of 80 people) as correct answers and evaluated whether the assigned headings are identical as the answers using five measures, including correctness (number of correct headings assigned / number of people).

The best pattern for the whole group of 80 people was "(a) the top ten pages, (b) the 100 characters before and after the person's name (total 200 characters), (c) a half weight for synonyms, and (d) df/N". The pattern's correctness was 26.3% (21/80).

We call this pattern the "highest best" and use it as the starting point of our research.

B. Assigning multiple headings

To identify people, multiple subject headings must be assigned to them. In this paper, we present methods that assign multiple NDLSHs to people on the web.

First, we investigated the highest best pattern. We manually evaluated the "relatedness" between the assigned headings and people by five values: 5: very related; 4: slightly; 3: neutral; 2: not very ; 1: unrelated.

Table I shows examples of the top ten results. Katsumi Tanaka 00 is a database researcher, Seiko Hishinuma 00 is a veterinarian in a fictional manga, and Susumu Goto 00 is a bioinformatics researcher. Both appropriate and inappropriate headings are found in the results. Veterinary medicine (5th rank) for Seiko Hishinuma 00 and bioinformatics (3rd rank) for Susumu Goto 00 are examples of such appropriate and correct answers.

We scrutinized the headings that both showed relatedness scores of 1 and appeared more than once; consequently we found that the major reasons for the occurrence was the use of synonyms. For example, "English (people)" was extracted since "English (language)" in web pages was selected as a heading "English (people)" using synonyms.

As a result of our observations, we adopted a pattern that stopped using synonyms from the highest best pattern, that is, the "(a) top ten pages, (b) 100 characters before and after the person's name (a total of 200 characters), (c) without synonyms, and (d) df/N." We call this pattern "no synonyms."

Table II shows examples of the top ten results of the no-synonym pattern. In it, we eliminated the inappropriate headings caused by using synonyms and reduced the number of inappropriate headings with value 1. Most unintelligible strings consisting of katakana characters were deleted. For example, no terms have a value of 1 and no unintelligible strings consist of katakana characters in Table II.

TABLE I
HIGHEST BEST PATTERN

Katsumi Tanaka 00	Seiko Hishinuma 00	Susumu Goto 00
database [5]	animal [4]	chemistry [4]
university [4]	*hamu* [1]	university [4]
information processing [5]	*itou* [1]	**bioinformatics** [5]
academic organization [3]	doctor [4]	education [4]
engineering [4]	**veterinary medicine** [5]	medicine [4]
object [2]	public health [4]	life science [5]
tara [1]	histidine [4]	database [4]
ten [1]	pseudonym [2]	English [1]
edit [2]	time, place [2]	Japanese [2]
LAN [3]	skiing [3]	science [4]
average [3.0]	average [3.0]	average [3.7]

Note 1: (a) top ten pages, (b) 100 characters before and after person's name, (c) a half weight for synonyms, and (d) df/N.
Note 2: [] denotes relatedness values. Italics denotes unintelligible strings consisting of Japanese katakana characters.

TABLE II
NO-SYNONYM PATTERN

Katsumi Tanaka 00	Seiko Hishinuma 00	Susumu Goto 00
database [5]	animal [4]	chemistry [4]
university [4]	**veterinary medicine**[5]	university [4]
information processing [5]	public health [4]	education [4]
engineering [4]	skiing [3]	life science [5]
object [2]	company [3]	medicine [4]
technology [4]	constitution [3]	**bioinformatics** [5]
edit [2]	university [4]	database [4]
research institute [3]	student [4]	creature [3]
explosion [2]	novel [3]	life [3]
student [2]	doctor [4]	gene [4]
average [3.3]	average [3.7]	average [4.0]

Note 1: (a) top ten pages, (b) 100 characters before and after the person's name, (c) no synonyms, and (d) df/N.
Note 2: [] denotes relatedness values.

We evaluated two patterns, "highest best" and "no synonyms," by cumulative relatedness (the average relatedness values from the highest to the corresponding ranks) (Fig. 1). The no-synonym pattern improved from the highest rank to the 10th rank and its values exceeded 3.

When assigning multiple NDLSHs to people, the no-synonym pattern (the one with identical conditions other than synonyms and without using them) outperformed the highest best pattern. We allocated the top ten headings of the no-synonym pattern, based on a related work [3] that assigned the top five NDC numbers with cumulative relatedness exceeding 3 to people.

III. PROTOTYPE SYSTEM

We developed "People on the Web (PoW)", a prototype system for searching people on the web by visually exploring

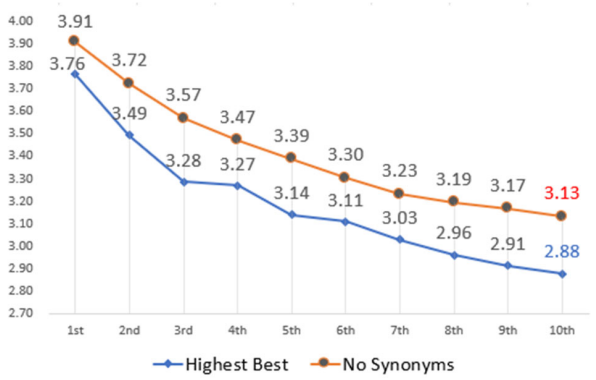

Fig. 1. Evaluation by cumulative relatedness

NDLSHs. People can be sought while exploring broader, narrower, and related heading terms. We assigned the top ten NDLSHs to 80 people in the dataset using the no-synonym pattern.

A. Facility

Entering a keyword in the search box produces a NDLSH list that includes the keyword. When the user selects a NDLSH, it appears in the main window. Double-clicking on the NDLSH displays a context menu. Selecting "NDLSH" displays broader, narrower, and related terms of the heading, and selecting "people" displays people (icons) to whom the heading is assigned. When users select a person icon, they can see the headings assigned to her in the right area. Double-clicking on a person icon and selecting "subject" from the context menu shows the NDLSHs assigned to him. The objects (NDLSH and people) can be moved, enlarged, reduced, etc.

B. Example of Use

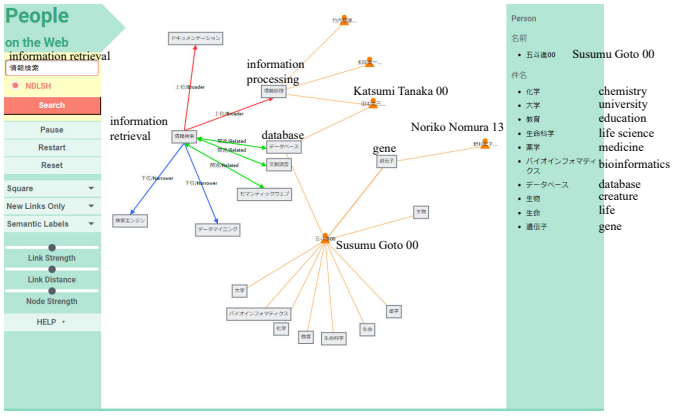

Fig. 2. Screen example

Figure 2 shows a use example. NDLSHs were searched in an exploratory manner from the heading "information retrieval." The "database" and "information processing" headings, which are related or broader terms of "information retrieval" and people (icons) to whom "database" or "information processing" are assigned, are displayed. When a user selects Susumu Goto 00, who is one of the persons to whom "database" is assigned, his assigned headings are displayed in the right area, and such detailed headings as "bioinformatics" can be checked. In addition, these headings, which are displayed from Susumu Goto 00 in the window and Noriko Nomura 13 to whom "gene" is assigned, will be displayed.

In this way, we can look for people by searching for related headings from the entered keywords or other persons through headings related to them.

IV. Related Work

To the best of our knowledge, no research assigns library subject headings to people on the web except our previous work (e.g. [1]). The methods for assigning controlled terms like subject headings to documents can be divided into with/without machine learning. Our research based on pattern-matching falls into the latter classification.

In our previous work, we assigned people location information [4], vocation-related information [5], and library classification numbers [3]. In this paper, we investigate a method that assigns NDLSHs to the results of web searches to help users select and identify people on the web. Our prototype system was developed by modifying our previous work [6], a system that visually explored subject headings.

V. Conclusions

We investigated methods of assigning multiple NDLSHs to people for searching and identifying them on the web. We developed a prototype system that visually explores NDLSHs and connected people through network representation. Future work will experiment with different datasets, explore various methods using machine-learning, and evaluate the effectiveness of our prototype system.

Acknowledgments: This work was supported by JSPS KAKENHI Grant Number 19K12718.

References

[1] M. Shimokura and H. Murakami, "Assigning NDLSH headings to people on the web," in *Information Retrieval Technology*, T. Tseng, Yuen-Hsienand Sakai, J. Jiang, L.-W. Ku, D. H. Park, J.-F. Yeh, L.-C. Yu, L.-H. Lee, and Z.-H. Chen, Eds. Springer International Publishing, 2018, pp. 189–195, doi: 10.1007/978-3-030-03520-4_18.

[2] S. Sato, K. Kazama, K. Fukuda, and K. Murakami, "Distinguishing between people on the web with the same first and last name by real-world oriented web mining," *IPSJ Transactions on Databases*, vol. 46, no. 8, pp. 26–36, 2005.

[3] H. Murakami, Y. Ura, and Y. Kataoka, "Assigning library classification numbers to people on the web," in *Information Retrieval Technology*, R. E. Banchs, F. Silvestri, T.-Y. Liu, M. Zhang, S. Gao, and J. Lang, Eds. Springer Berlin Heidelberg, 2013, pp. 464–475, doi: 10.1007/978-3-642-45068-6_40.

[4] H. Murakami, Y. Takamori, H. Ueda, and S. Tatsumi, "Assigning location information to display individuals on a map for web people search results," in *Information Retrieval Technology*, G. G. Lee, D. Song, C.-Y. Lin, A. Aizawa, K. Kuriyama, M. Yoshioka, and T. Sakai, Eds. Springer Berlin Heidelberg, 2009, pp. 26–37, doi: 10.1007/978-3-642-04769-5_3.

[5] H. Ueda, H. Murakami, and S. Tatsumi, "Assigning vocation-related information to person clusters for web people search results," in *2009 WRI Global Congress on Intelligent Systems*, vol. 4, 2009, pp. 248–253, doi: 10.1109/GCIS.2009.254.

[6] H. Murakami, "A system for exploring NDLSH and LCSH headings," in *IFLA WLIC 2019*, 2019.