Assigning NDLSH Headings to People on the Web

Masayuki Shimokura and Harumi Murakami

Graduate School for Creative Cities, Osaka City University,
3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585, Japan
shimokura@gmail.com

Abstract. We investigate a method that assigns National Diet Library Subject Headings (NDLSH) to the results of web people searches to help users select and understand people on the web. NDLSH is a controlled subject vocabulary list compiled and maintained by the National Diet Library (NDL) as a subject access tool. By assigning NDLSH headings to people, well-formed keywords can be assigned, and exploratory searches using related terms are possible. We examined the following combination of factors: (a) web-page rank (the number of pages), (b) position inside the HTML, (c) synonyms, and (d) document frequency. We report our experimental results for 405 combination patterns ($5 \times 9 \times 3 \times 3$) using our 80-person dataset. Overall, under our experimental settings, the best combination was (a) the top ten pages, (b) 100 characters before and after a person's name (i.e., 200 characters), (c) half weight for synonyms, and (d) document frequency divided by number of web pages.

Keywords: web people search, subject headings assignment, NDLSH, experiment

1 Introduction

The popularity of web people searches continues to rise as the number of people increases about whom the web can provide information. If the list of web people search results 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.

In our previous work [1], we assigned people location information [2], vocation-related information [3], and library classification numbers [4]. In this paper, we investigate a method that assigns National Diet Library Subject Headings (NDLSH) to the results of web people searches to help users select and understand people on the web. The NDLSH [5] is a controlled subject vocabulary list compiled and maintained by the National Diet Library (NDL) as a subject access tool. By assigning NDLSH headings to people, well-formed keywords can be assigned, and exploratory searches using related terms are possible.

We examined the following combination of factors: (a) web-page rank (the number of pages), (b) position inside the HTML, (c) synonyms, and (d) document frequency. We report our experimental results for 405 combination patterns ($5 \times 9 \times 3 \times 3$) using our 80-person dataset.

Below, we explain our experiment in Sections 2 and 3, and the significance of our research in Section 4. The examples presented in this paper were translated from Japanese into English for publication.

2 Method

2.1 Procedure

As queries, we used 20 Japanese names from a related work [6] and obtained 50 web pages per each query via Google Custom Search API. We manually classified these pages into different people and identified 80 separate people. NDLSH headings were assigned to HTML files for each person.

First, we extract headings with variants (synonyms) and delete those with two or fewer single-byte alphanumeric characters, those with only one double-byte character, and those contain "--(two hyphens)" because they are less important terms and/or not very useful for text-matching.

With regard to terms (headings and synonyms), since longer strings provide more specific meaning, we count the number of terms (headings and synonyms) from the longer ones in the HTML documents without tags in 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.

We prepared the following four types of combination conditions:

(a) Search ranking of web pages: five patterns of top 1, 3, 5, 10, and all pages for each person.
(b) Position 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 that don't use synonyms, using synonyms with identical weights as the headings, and using synonyms with half of the weight of the headings.
(d) Document frequency of headings and synonyms: three patterns of doing nothing, multiplying document frequency (df) / total number of used documents (N) (i.e., multiplying df/N), and multiplying total number of used documents (N) / document frequency (df) (i.e., multiplying N/df). The last one is rephrased as multiplying inverse document frequency (i.e., multiplying idf).

When these conditions are combined, they become $5 \times 9 \times 3 \times 3 = 405$ patterns. Fig.1 shows an example of the score calculation for headings. A heading having with the highest score is assigned to the corresponding person. If no heading is assigned, the answer becomes "none".

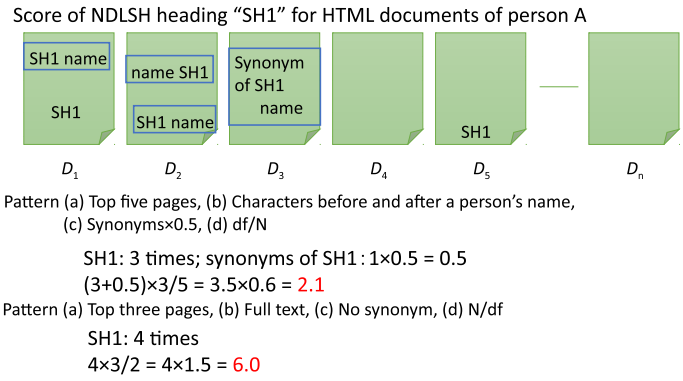

Fig. 1. Example of score calculation

2.2 Evaluation

We manually selected the most appropriate NDLSH heading for each person (79 out of 80 people). For example, a "baseball" heading was selected for Suguru Egawa, a former professional baseball player, and "social psychology" was selected for Asako Miura, a Kwansei Gakuin University professor.

The following are the evaluation measures:

$$\text{Correctness} = \frac{\text{number of correct NDLSHs assigned automatically}}{\text{number of people}}$$

$$\text{Precision} = \frac{\text{number of correct NDLSHs assigned automatically}}{\text{number of people to whom an NDLSH was assigned automatically}}$$

$$\text{Recall} = \frac{\text{number of correct NDLSHs assigned automatically}}{\text{number of people to whom an NDLSH was assigned manually}}$$

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

$$\text{Accuracy} = \frac{\text{number of correct answers}}{\text{number of people}}$$

When calculating the Accuracy, none is judged correct when there is no correct NDLSH heading for a person.

3 Results and Analysis

Table 1 shows which pattern had the best correctness. The best pattern for the whole (80 people) was "top ten pages, 100 characters before and after the person's name (total 200 characters), 0.5 times for synonyms, and df/N," and its correctness was 26.3% (21/80).

Table 1. Patterns of high correctness

Number of Documents	Ranking	Position	Synonym	Document Frequency	Correctness
All	10	Before and After 100 Characters	0.5	df/N	0.263 (21/80)
1	1	Full Text	None	1	0.286 (10/35)
2	3	Before and After 200 Characters	0.5	1	0.333 (4/12)
3 or more	10	Before and After 100 Characters	1	df/N	0.364 (12/33)
11 or more	10	Before and After 100 Characters	1	df/N	0.500 (9/18)
3 to 10	5	Before and After 60 Characters	0.5	df/N	0.267 (4/15)

Since we observed that the trend may differ depending on the number of documents for each person, we classified the number of documents into 1, 2, 3 or more, 11 or more, 3 to 10, and conducted our evaluation. For 35 people with only one document, "full text (not using synonyms)" was best. For 33 people with 3 or more documents, the result was almost the same as for all the people (the difference is only the magnification of synonyms), and the result was the same for 18 people with 11 documents or more.

Table 2. Evaluation for patterns of high correctness

Number of Documents	Precision	Recall	F-measure	Accuracy
All	0.276	0.266	0.271	0.263
1	0.286	0.294	0.290	0.286
2	0.333	0.333	0.333	0.333
3 or more	0.364	0.364	0.364	0.364
11 or more	0.500	0.500	0.500	0.500
3 to 10	0.182	0.267	0.216	0.267

Evaluations for high correctness patterns are shown in Table 2. Precision and recall are less than 30% for both all people and people with one document; they are 50% for people with 11 or more documents.

Table 3. Evaluation using cosine

Number of Documents	Precision	Recall	F-measure	Accuracy
All	0.026	0.025	0.026	0.025

For comparison, we implemented a method using cosine as a baseline. MeCab [7], a Japanese morphological analyzer, was used to extract terms from documents (web pages), headings and synonyms; nouns that contain two or more characters were extracted as terms. To match the conditions to our best pattern, we used the top ten pages, 100 characters before and after a person's name (200 characters), synonyms, and document frequency. The weights of the terms extracted from the headings were normalized by frequency $\times$ df/N, and the weights of the terms extracted from the synonyms were 0.5 times normalized frequency $\times$ df/N. Table 3 shows the evaluation results with cosine. Our best pattern significantly outperformed cosine.

Based on the above results, we found that we must narrow down to the top ten rather than all of the whole web pages, use the character strings before and after a person's name rather than the full text of the HTML documents, and weight the words that appear in many documents using synonyms. However, the best approach was counting the number of headings from the full text for those with only one document. These performances were much better than the method using cosine.

4 Related Work and Discussion

We roughly divided the methods for assigning controlled terms to documents into with and without machine learning. Our research falls in the latter classification. Cosine is one of the most commonly used baselines for non-machine learning methods. The results obtained in this research are significantly higher than the baseline.

In research that assigns Nippon Decimal Classification (NDC) numbers to people on the web to develop a person directory [4], we considered two conditions of documents $\times$ methods. The title (texts inside title tags) was best among the document conditions, and the result differs from this research. We believe that this is due to the difference of research purposes and dataset. [8] suggested that the best patterns were different based on the number of documents per person (i.e., 1, 2, 3 or more documents), and this is related to this research.

There is research that assigns labels to people except for our previous work. Wan et al. [9] assigned titles (including vocations) and Mori et al. [10] assigned keywords to person clusters. WePS-2/3 [11] conducted competitive evaluation on person attribute extraction on web pages. No such research has assigned subject headings to people on the web.

Our experimental results revealed the best pattern for assigning NDLSH headings to people on the web. To the best of our knowledge, this is the first research that assigns library subject headings to people on the web.

5 Conclusions

We investigated a method that assigns NDLSH headings to the results of web people searches to help users select and understand people on the web. We exam-

ined the following combination of factors: (a) web-page rank, (b) position inside HTML, (c) synonyms, and (d) document frequency. We reported the results of our experiment for 405 patterns with an 80-person dataset. Overall, under our experimental settings, the best combination was (a) the top ten pages, (b) 100 characters before and after a person's name (i.e., 200 characters), (c) half weight for synonyms, and (d) document frequency divided by the number of web pages.

Future work will examine the use of stop lists and such terms as broader, narrower, and related terms. We also need to evaluate our method for other headings (except for highest score) and with large and diverse datasets.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 25330385, 16K00440.

References

1. Murakami, H., Ueda, H., Kataoka, S., Takamori, Y., Tatsumi, S.: Summarizing and Visualizing Web People Search Results. In: Proceedings of the Second International Conference on Agents and Artificial Intelligence (ICAART 2010), vol. 1, pp. 640–643. INSTICC Press (2010)
2. Murakami, H., Takamori, Y., Ueda, H., Tatsumi, S.: Assigning Location Information to Display Individuals on a Map for Web People Search Results. In: Lee, G.G., 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)
3. 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)
4. Murakami, H., Ura, Y., Kataoka, Y.: Assigning Library Classification Numbers to People on the Web, In: Banchs et al. (eds.), AIRS 2013, LNCS, vol. 8281, pp.464–475, Springer, Heidelberg (2013)
5. Cataloging Tools and Resources, http://www.ndl.go.jp/en/data/classification_subject.html
6. Sato, S., Kazama, K., Fukuda, K., Murakami, 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. MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/
8. Murakami, H., Ura, Y., Kataoka, Y.: Assigning Library Classification Numbers to People on the Web and Developing People-search Directory, Transactions of the Institute of Systems, Control and Information Engineers, 29(2), 51-64 (2016)(in Japanese)
9. 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)
10. Mori, J., Matsuo, Y., Ishizuka, M.: Personal Keyword Extraction from the Web. Journal of Japanese Society for Artificial Intelligence 20, 337-345 (2005)
11. 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)