Suggesting Japanese Subject Headings using Web Information Resources

Hiroshi Ueda
Graduate School of Engineering, Osaka City University
d06tb001@ex.media.osaka-cu.ac.jp
3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 Japan

Harumi Murakami
Graduate School for Creative Cities, Osaka City University
harumi@media.osaka-cu.ac.jp
3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 Japan

Introduction

It is well-known that suggesting subject headings according to user queries is useful for subject searches (e.g., the MeSH database for MEDLINE ("MeSH databse," n.d.), RedLightGreen ("RedLightGreen," n.d.)). However, few systems suggest subject headings for OPACs in Japan. Some systems (e.g. "National Diet Library NDL-OPAC," n.d.), which offer users a function that retrieves subject headings, are based on pattern matching algorithms, and they usually fail to suggest a subject heading when there is no hit.

In this paper, we propose a method that suggests BSH4 (Japan Library Association, 1999) subject headings according to user queries when pattern matching algorithms fail to produce a hit. As user queries are diverse and unpredictable, we explore a method that makes a suggestion even when the query is a new word. We investigate the use of information obtained from Wikipedia ("Wikipedia," n.d.), the Amazon Web Service (AWS), and Google. We implemented the method, and our system suggests ten BSH4 subject headings according to user queries.

Method

Overview

User queries are expanded using information obtained from Wikipedia, AWS, and Google. BSH4 subject headings are also expanded using narrower terms. The similarity between the query and subject headings is calculated using a cosine measure, based on a vector space model (Grossman & Frieder, 2004).

Query Expansion using Web Information Resources

Using Wikipedia, AWS, and Google

Wikipedia is an encyclopedia that can be freely used on the Web. According to Guiles (2005), Wikipedia comes close to Britannica in terms of the accuracy of its science entries. We think it is beneficial to use information obtained from Wikipedia to get new definitions and information.

AWS is a service for developers who offer various data concerning goods, books, and other products in Amazon. BrowseNodes tags express category data for books. We use the texts of BrowseNodes tags that are dispatched to books because we think the category data for books resemble subject headings.

Google is one of the most frequently used Web search engines. When we cannot get information from Wikipedia or AWS, Google functions as a complementary resource for query expansion.

Collecting information from the Web

The system first searches Wikipedia using the query through its top page. If there is one result, then it gets it, otherwise, it searches the Google site search for Wikipedia. If there are results, then it gets the 1st result.

It then searches AWS, and if there are results, then it gets the 1st to 3rd results.

Finally it searches Google; if there are results, then it gets the 1st to 5th results.

Creating a query vector

To identify terms for creating vectors, nouns consisting of two or more characters are extracted from collected information using morphological analysis. This process is for Japanese characters.

Terms extracted from Wikipedia are weighted three times.

Example of query expansion

When "Java" was input as a query, a query vector becomes [object (30), oriented (30), programming (29), language (27), Java (8), software (2), internet (1), computer (1), download (1), general (1), free (1) …and so on].

Examples in this paper were originally Japanese. They were translated into English for publications.

Expanding subject headings using narrower terms

Narrower terms are included to expand document vectors. Nouns consisting of two or more characters were extracted using morphological analysis as the query vector.

For example, subject heading "information retrieval" has a top term "information science," a narrower term "indexing method," and "database." The "indexing method" has a narrower term "punch card." In this case, a document vector for "information retrieval" becomes [information (1), retrieval (1), indexing (1), "punch (1), data (1), and "base (1)"].

Example

When "Java" was input as a query, the suggested ten subject headings were "computer programming," "programming (computer)," "internet," "computer graphics," "computer art," "computer music," and "computer crime", "personal computer", "computer network", and "Kanji processing (computer)" on July 10th, 2005.

Experiment

We used computer terms ("e-Words," n.d.) as a query.

Experiment 1

Method

The subjects were 41 undergraduate students. 99 computer terms were used for the experiment. A questionnaire sheet displayed one computer term and ten subject headings suggested by our method. Five questionnaires were allocated to the subjects.

First, the subjects evaluated their familiarity with the computer term: (5: very well; 4: well; 3: neutral; 2: not very well; 1: not at all). We call this the "known" measure.

Next, they evaluated how much these suggested subject headings were related to the associated computer term (3: related; 2: neutral; 1: unrelated.). We call this the "related" measure.

We judged data whose related value is "3" as relevant to calculate "precision."

The experiment was conducted on July 14th, 2005.

Results and Discussion

The subjects rarely filled related values to subject headings when they answered "1" or "2" as known values for the computer term. Therefore, we analyzed the data whose known values were more than 2.

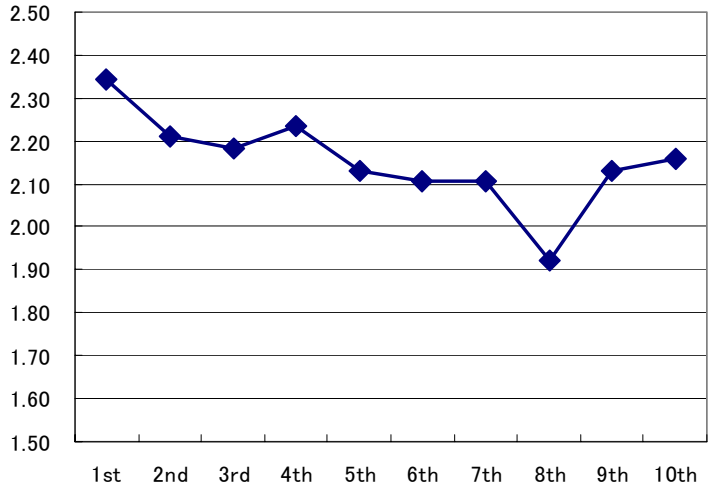

Figure 1. Average related values at each rank

Figure 1 shows the average related values at each rank of ten suggested subject headings. The related value of the top rank (1st) was the highest (2.34).

Precision was 55% at the top rank (1st), 49% at top 3, and 41% at top 10.

The above results suggest the usefulness of our method when users input known terms as queries.

Experiment 2

Method

We examined the effectiveness of a combination of Wikipedia, AWS, and Google.

We compared our method (d) with each resource: (a) Wikipedia, (b) AWS, and (c) Google. We used 33 computer terms. Questionnaire sheets were identical to Experiment 1.

After the subject, a computer science graduate student, read the definitions of 33 computer terms, she evaluated how much the suggested subject headings were related to the computer terms, as in Experiment 1. She filled 33 (terms) * and 4 (conditions) questionnaire sheets.

The experiment was conducted on Dec 7th, 2005.

Results and Discussion

For the average related value of the top rank (1st), our method (d) was highest (2.35) compared to (a) Wikipedia (1.90), (b) AWS (2.19), and (c) Google (1.90).

Table 1. Precision for four conditions

	Top 1	**Top 3**	**Top 10**
(a) Wikipedia	39%	25%	24%
(b) AWS	58%	47%	38%
(c) Google	35%	27%	19%
(d) our method	58%	55%	43%

Table 1 outlines the precision results for each condition at top 1, top 3, and top 10. Precision was 58% at the top rank, 55% at top 3, and 43% at top 10. Our method was the highest among all conditions.

We found that the combination of Wikipedia, AWS, and Google was useful for query expansion.

Summary

We proposed a method that suggests subject headings according to user queries for subject searches for OPACs. We combined information obtained from Wikipedia, AWS, and Google for query expansion.

Two experimental results suggest that (1) our method is useful when users input known terms, and (2) the combination of Wikipedia, AWS, and Google is useful for suggesting subject headings.

References

e-Words noteworthy term ranking 100. (n.d.). Retrieved June 5, 2006, from http://e-words.jp/p/s-ranking.html

Grossman, D. A., & Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics* (2nd ed.). Springer.

Japan Library Association. (1999). B.S.H. Basic Subject Headings 4th edition.

MeSH Database. (n.d.). Retrieved June 5, 2006, from http://www.ncbi.nlm.nih.gov/entrez/

National Diet Library NDL-OPAC. (n.d.). Retrieved June 5, 2006, from http://opac.ndl.go.jp/

RedLightGreen. (n.d.). Retrieved June 5, 2006, from http://www.redlightgreen.com/

Wikipedia. (n.d.). Retievied June 5, 2006, from http://ja.wikipedia.org