Automated Subject Induction from Query Keywords through Wikipedia Categories and Subject Headings

Yoji Kiyota[§], Noriyuki Tamura[¶], Satoshi Sakai[¶], Hiroshi Nakagawa[§], Hidetaka Masuda[¶]

§Information Technology Center, University of Tokyo
General Library, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 Japan
E-mail: kiyota@r.dl.itc.u-tokyo.ac.jp, n3@dl.itc.u-tokyo.ac.jp

¶Graduate School of Engineering, Tokyo Denki University
2-1 Kanda-nishiki-cho, Chiyoda-ku, Tokyo 101-0054 Japan
E-mail: {tamura,sakai}@cdl.im.dendai.ac.jp, masuda@im.dendai.ac.jp

Abstract

This paper addresses a novel approach that integrates two different types of information resources: the World Wide Web and libraries. This approach is based on a hypothesis: advantages and disadvantages of the Web and libraries are complemental. The integration is based on correspondent conceptual label names between the Wikipedia categories and subject headings of library materials. The method enables us to find locations of bookshelves in a library easily, using any query keywords. Any keywords which are registered as Wikipedia items are acceptable. The advantages of the method are: the integrative approach makes subject access of library resources have broader coverage than an approach which only uses subject headings; and the approach navigates us to reliable information resources. We implemented the proposed method into an application system, and are now operating the system at several university libraries in Japan. We are planning to evaluate the method based on the query logs collected by the system.

1. Intoduction

During the last decade, the principal method of information retrieval for people is drastically shifted: from libraries to web search engines. Using web search engines, we can hit a lot of web pages which are related to our interest. Information on the Web helps us to solve our problems in most cases, however, the Web have shortcomings. First, the information is not always well organized. If we inputs a vague query keyword (e.g., earthquake), a huge number of pages are simply listed. While any types of web pages related to the query (e.g., hazard statistics, news, predictions, and fictions) are jumbled together in the list, we are often confused. Second, a significant part of information on the Web is unreliable. When we want to verify reliability of a web page, we often have to refer to other information re-sources such as news archives and journals in libraries.

Meanwhile, in response to changes on the Web, new trends in libraries are observed. For example, some librarians began to utilize some information on the Web, typically Wikipedia, for providing reference services. Web sites of some libraries have pathfinders [Cohen 1999], each of which provides useful information for people who begin to retrieve information on a specific topic, such as introduction, the Dewey Decimal System (DDC) code, reference books, and useful web sites related to earthquake. However, such useful services provided by libraries do not play a principal role of information retrieval, because these services require huge human resources. Due to limited budget, libraries cannot keep up with people's demand now.

As stated above, information retrieval which relies on a single type of information resource, both the web only and libraries only, has limitations respectively. Our solution to the limitations is integration of the taxonomies each of which represents each information resource: subject headings of libraries and Wikipedia categories.

2. Material Organization Systems of libraries and Wikipedia categories

This section shows a brief comparison between material organization systems of libraries and Wikipedia, focusing on those advantages and shortcomings.

2.1 Material Organization Systems of Libraries

Each material in libraries (e.g. books, serials, maps, and multimedia materials) is organized based on organization systems, according to their subjects. Usually, the organization systems group entities that are similar together arranged in a hierarchical tree structure.

Material organization systems chiefly consist of two types of tools. One is library classification systems, which allocate a call number to each material. Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) are widely used. The other is subject headings systems, which assign keywords to each material. The Library of Congress Subject Headings (LCSH) are widely used.

These organization systems have consistent hierarchical structures, and are usually maintained by professionals in the field of library and information science. Revision of the systems requires careful judgment so that the hierarchical structures keep consistency. Usually, the systems are managed through committee-based approach. The advantages of the organization systems of libraries are:

- **Stability.** Most of the systems have semi-permanent structures. Even if classifications and subject headings have to be modified, compatibility of the systems is maximally taken into account. The stability enables us to make use of the systems

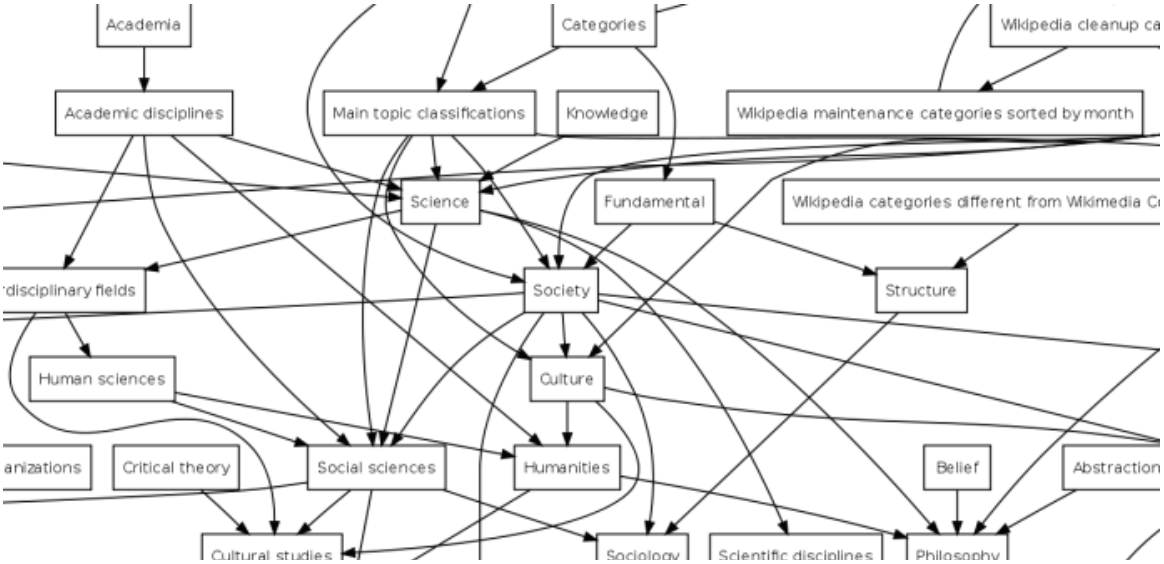

Figure 1: A snapshot of the structure of Wikipedia categories.

comfortably.

- **Deep organization structure.** The structure gives us an overhead view of the domain we are interested in.

The shortcomings of the systems are:

- **Lack of newly-appeared concepts.** Due to the committee-based approach, introduction of new concepts tends to be late.
- **Lack of diversity.** Each concept is usually given only one broader concept in the hierarchical structure, so the various aspects tend to be ignored.

2.2 Wikipedia Categories

Wikipedia is a multilingual, web-based, encyclopaedia project operated by the Wikimedia Foundation. As of September 2007, Wikipedia had approximately 8.29 million articles in 253 languages, comprising a combined total of over 1.41 billion words for all Wikipedias.

As various people over the world participate in editing articles, Wikipedia potentially covers almost all concepts in the world. In addition, Wikipedia organizes enormous number of articles into categories. The Wikipedia categorization system was originally designed to browse through similar articles. The categorization system is regarded as a folksonomy system, because any editors can assign any free tags (categories) to each Wikipedia item. For example, the item "Price" has multiple tags such as "Marketing", "Market", and "Economics". Furthermore, categories also have broader categories. As a result, the categorization system forms a relaxed style of hierarchical structure (Figure 1).

Since the Wikipedia category structure is built based on a bottom-up approach, the structure has the following advantages:

- **Quick introduction of newly-appeared concepts.** Without any restrictions for using new category names, the number of category names is increasing rapidly.
- **Flexibility.** Since the number of categories for each Wikipedia item and category is not limited, the assigned categories can reflect various aspects of the concept.

On the other hand, the Wikipedia category structure has the following shortcomings which come from the bottom-up approach:

- **Lack of stability.** Since any people can edit the Wikipedia, the category structure is changing rapidly. So navigation using the structure is not always reliable.
- **Shallow organization structure.** Some Wikipedia items and categories are not well organized, because they do not have appropriate categories

3. Solution: Integration of Wikipedia Categories and Subject Headings

As mentioned above, both material organization systems of libraries and Wikipedia categories have advantages and shortcomings. Since the advantages and the shortcomings are complementary, use of the both structures as situations is needed. This paper proposes a method that realizes complementary use of the both structures: beginning up from a Wikipedia item, and inducing related subject headings via the structure of Wikipedia categories.

Figure 2 shows the overview of our method. Firstly, we describe use of Wikipedia as a start point of information retrieval. Suppose that we begin retrieval from a keyword "Hanshin Great Earthquake". In the Japanese version of Wikipedia, the Wikipedia entry "Hanshin Great Earthquake" has categories such as "History of earthquakes" and "Economic history of Japan". The category "History of earthquakes" also has broader categories such as "History of hazards" and "Earthquake", and the category "Economic history of Japan" has a broader category "Economic history". As a result, we can get a subset of related categories as a tree structure. This tree structure seems to be a cross-section of "Hanshin Great Earthquake". For example, the path "Economic history of Japan" "Economic history" shows that the

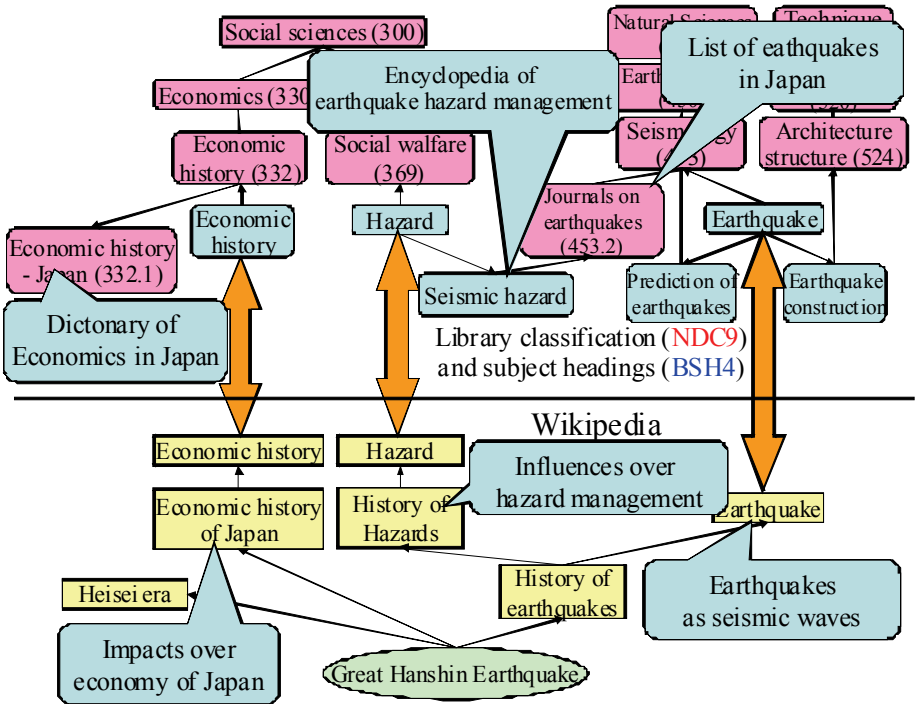

Figure 2: Induction of subject headings via the networks of Wikipedia categories.

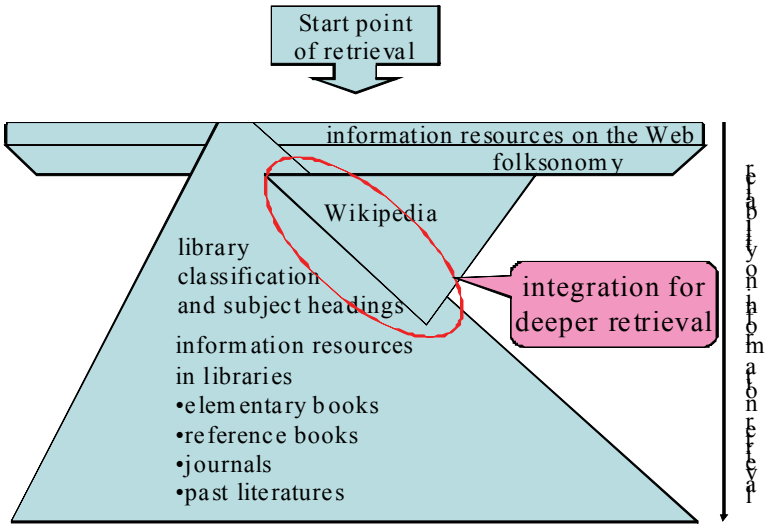

Figure 3: Navigation toward reliable information resources from any query keywords.

earthquake can be views as “impacts over economy of Japan”, and the path “History of earthquakes” “History of hazards” “Hazard” shows that it can be views as “influence over hazard management in Japan”. As a result, the given subject “Hanshin Great Earthquake” can be generalized into “Economic history”, “Hazard”, and “Earthquake”.

Second, we describe correspondences between Wikipedia categories and subject headings, and application of subject headings for information retrieval. As described below, there are overlaps between Wikipedia categories and subject headings. In Figure 2, Wikipedia categories “Economic history”, “Hazard”, and “Earthquake” are correspondent with subject headings of BSH4 (Basic Subject Headings) [JLA 1999], which is developed by Japan Library Association. Each BSH4 subject heading is associated with NDC9 (Nippon Decimal Classification) [JLA 1995], which is a widely used classification system in Japanese libraries. Our approach applies the overlaps between Wikipedia categories and subject headings, to retrieve useful information resources in libraries. For example, following the path “Economic history” “332” “332.1 (Economic history - Japan)”, we can find a reference book “Dictionary of Economics in Japan”.

We think that the integrative approach is useful because of the following reasons.

Figure 4: The screenshot of "Littel Navigator".

- **Sufficient overlaps.** We found that there are a lot of subject headings which have correspondence with Wikipedia categories. For example, out of approx. 11,000 subject headings in BSH4, there are approx. 1,400 subject headings which correspond to categories in Japanese version of Wikipedia. Note that there are approx. 15,000 categories in Wikipedia.
- **Broad coverage of concepts.** Our method can overcome the shortcomings of the low coverage of subject headings, by extending it with Wikipedia. Since Wikipedia potentially covers almost all concepts in the world, the method will be universal.
- **Navigation toward reliable information resources.** If we expand a query keyword using only Wikipedia, the induction of categories will not be useful, because they are not necessarily associated with reliable information resources. Giving subject headings of libraries, the expansion is useful for reliable information retrieval. Figure 3 shows the usefulness of our approach.

4. An Application System

To evaluate the effectiveness of our method, application to real situations of information retrieval is inevitable. We have been paying for digital reference services (DRS) of libraries, because DRS is a good choice for collecting large amounts of information queries of whom need reliable information resources. We are now attempting to apply the proposed method to information retrieval in libraries. Specifically, we developed a prototype system "Littel Navigator", and operate the system in several university libraries in Japan.

Figure 4 shows the screenshot of Littel Navigator. If you inputs query keywords (e.g., Hanshin-Awaji Great Earthquake), the system outputs induced "themes" related to keywords such as "earthquake", "economy", "seismology" and "disaster", in addition to reliable information resources, including a reference book summarizing the history of great earthquakes in Japan.

We are planning to estimate the usefulness of the method, by evaluating the operation logs of Littel Navigator.

5. Conclusion

This paper addresses potentials of a new infrastructure for information retrieval, which integrates two types of information resources: the Web and libraries. The integration will bridge two paradigms of classification: taxonomy and folksonomy. Natural language processing techniques, including similarity calculation and acquisition synonyms, will contribute to enhancement of the potentials.

6. References

Laura B. Cohen and Julie M. Still. (1999). A comparison of research university and two-year college library web sites: content, functionality, and form, *Collage and research libraries*, Vol. 60, No. 3, pp. 275-289, 1999.

Japan Library Association (ed.). (1999). Kihon Kenmei Hyoumokuhyou (Basic Subject Headings) 4th edition.

Kiyoshi Mori, Japan Library Association (ed.). (1995). Nippon Decimal Classification 9th edition.