

Exploitation of the Wikipedia Category System for Enhancing the Value of LCSH

Yoji Kiyota
Information Technology Center
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-0033 Japan
+81-3-5841-2738
kiyota@r.dl.itc.u-tokyo.ac.jp

Satoshi Sakai
Graduate School of Engineering
Tokyo Denki University
2-1 Kanda-nishiki-cho, Chiyoda-ku
Tokyo 101-0054 Japan
+81-3-5280-3551
sakai@cdl.im.dendai.ac.jp

Tatsuya Mori
Graduate School of Engineering
Tokyo Denki University
2-1 Kanda-nishiki-cho, Chiyoda-ku
Tokyo 101-0054 Japan
+81-3-5280-3551
mori@cdl.im.dendai.ac.jp

Hidetaka Masuda
Graduate School of Engineering
Tokyo Denki University
2-1 Kanda-nishiki-cho, Chiyoda-ku
Tokyo 101-0054 Japan
+81-3-5280-3551
masuda@im.dendai.ac.jp

Hiroshi Nakagawa
Information Technology Center
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-0033 Japan
+81-3-5841-2738
n3@dl.itc.u-tokyo.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 complemented. 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.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval. Content Analysis and Indexing; K.4.3 [Computers and Society]: Organizational Impacts—Computer supported collaborative work

General Terms

Design, Algorithms, Management, Experimentation

Keywords

Subject headings, Wikipedia Categories, LCSH, hybrid classification,

1. INTRODUCTION

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 classification systems each of which represents each information resource: subject headings of libraries and Wikipedia categories.

This paper consist of the following sections: Section 2 shows a brief overview of previous work on integrating Web resources and library resources. Section 3 pointed out the unique feature of Wikipedia categories, based on discussion on the differences between two classification paradigms, that is, top-down classification mainly applied to library resources, and bottom-up classification mainly applied to Web resources. After that, Section 4 proposes a solution to exploit advantages of both top-down classification and bottom-up classification. The solution integrates subject headings (i.e. LCSH: Library of Congress Subject Headings) and the Wikipedia categories, based on a score propagation algorithm for the Wikipedia category system. Section 5 shows some experiments on the proposed solution, and suggests the usefulness of the solution. Section 6 describes the overview of Littel Navigator, a navigation system for supporting reference services, as an application of the solution. Finally, Section 7 discusses the advantages of the proposed solution, and concludes this paper.

2. INTEGRATION OF WEB RESOURCES AND LIBRARY RESOURCES

As stated above, Web search engines have several advantages on information retrieval, and have been used by a lot of ordinary people, however, those also have disadvantages on credibility and quality of the outputs. In contrast, systems of libraries (e.g. OPAC, online databases and websites of libraries) have advantages on credibility and quality, but have disadvantages on usability and coverage of queries. In this section, we show a brief overview on previous work on integrating advantages of both Web resources and library resources, based on differences of philosophies between the Web world and the library world.

Several studies indicated that the appearance of Web search engines had a major impact on peoples' search behaviors. OCLC Online Computer Center published a survey report [7] on the attitudes toward information resources, targeting on students of six countries. According to the report, when asked which resource they turn first when they are looking for information, 89 per cent of the students indicated Web search engines, while only 2 per cent indicated online databases, and another 2 per cent indicated their library's website. Sadeh [5] indicated that Web search engines have:

- the ease of use, ease of access, and speed that characterize internet tools and services,
- the availability of integrated search environments,
- the new internet tools and services place on the user in adherence to the Web 2.0 design concepts, and
- cyber-interaction: users consider the internet a meeting place that enables them to exchange scholarly and non-scholar information.

On the other hand, the shortcomings of systems of libraries have been indicated. Sadeh told that "currently library systems are

typically disconnected from user spaces and expect some effort on the part of the user to access library collections". Furthermore, the high cost on cataloging has become a big issue in libraries. The report by Karen Calhoun [2] indicated "Over time, catalog construction has been constrained by the high cost of cataloging, the small size of a catalog card, and the scarcity of full text". The report also argued that the need of considering the balance between coverage and quality.

However, several studies indicated that the role of libraries remain important. The OCLC report [6] shows that most of the students (77 per cent) believed that the library resources are trustworthy or credible, and 75 per cent agree that librarians add value to the information search engines. Sadeh pointed out that libraries has the following roles:

- libraries offer quality information resources that librarians have carefully selected to meet their users' needs
- Not only can a library designate a spectrum of authoritative scholarly information, but it can also offer slices of such information to individual users on the basis of their affiliation and personal preference
- libraries can offer a clear statement of an item's availability and the means for obtaining it

These roles can be realized only if the materials of libraries are appropriately classified.

Lois Mai Chan [8] indicated that seamless integration of Web resources and library resources requires the following conditions:

- interoperability among different systems, metadata standards, and languages
- flexibility and adaptability to different information communities, not only different types of libraries, but also other communities such as museums, archives, corporate information systems, etc
- extensibility and scalability to accommodate the need for different degrees of depth and different subject domains
- simplicity in application, i.e., easy to use and to comprehend
- versatility, i.e., the ability to perform different functions
- amenability to computer application

We think that reexamination of the shape of classification systems (LCSH, LCC and so on) to realize the integration.

3. CLASSIFICATION PARADIGMS: TOP-DOWN, BOTTOM-UP, AND HYBRID APPROACHES

As indicated above, the way of the organization systems of libraries (LCSH, LCC and so on) should be reexamined. This section pointed out the unique feature of Wikipedia categories, based on discussion on the differences between two classification paradigms, that is, top-down classification mainly applied to library resources, and bottom-up classification mainly applied to Web resources.

3.1 Top-down Classification

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 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.

3.2 Bottom-up Classification

Recently, so-called folksonomy systems have been widely used by popular web services, including YouTube, del.icio.us, and Flickr. The folksonomy system can be regarded as a bottom-up classification. This approach 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 tags for each item is not limited, the assigned tags can reflect various aspects of the concept.

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

- **Lack of stability.** Since any people can edit the tags, the tag structure is changing rapidly. So navigation using the structure is not always reliable.
- **Shallow organization structure.** Since folksonomy tags are not well organized, because they do not have hypertexts.

3.3 Hybrid Classification: 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).

Voss [3] explored the category system of Wikipedia, and indicated that the system has a thesaurus-like structure that combines collaborative tagging (bottom-up approach) and hierarchical subject indexing (top-down approach) in a special way.

4. A SOLUTION: INDUCTION OF SUBJECT HEADINGS THROUGH WIKIPEDIA CATEGORIES

As mentioned above, the system of Wikipedia categories has unique characteristics, that is, a collaborative tagging system which has hierarchical structure. This paper proposes a method that integrates Wikipedia categories and subject headings: beginning up from a Wikipedia entry, and inducing related subject headings via the structure of Wikipedia categories.

4.1 Overview of Our Method

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 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), which is developed by Japan Library Association. Each BSH4 subject heading is associated with NDC9 (Nippon Decimal Classification), 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”.

4.2 Definition

In order to induce appropriate subject headings for any entry in Wikipedia, it is inevitable to define “*appropriateness* of subject headings”. If *appropriateness* is defined as a score calculation algorithm, it is useful to rank subject headings for any entry. We assume that each *appropriate* subject heading should satisfy the following two conditions:

- *The connection of the path* between the subject heading and the entry in Wikipedia is *stronger*. In other words, each link on the path should connect *similar* nodes (the entry, the subject heading, and categories on the path). In Figure 2, the entry “Hanshin Great Earthquake” and a subject heading “Hazard” are connected by three links: [“Hanshin Great Earthquake” - “History of earthquakes”], [“History of earthquakes” - “History of Hazards”], and [“History of Hazards” - “Hazard”]. These three links seem to connect *similar* nodes, that is, the connected two nodes of each link share common concepts. For example, the nodes of the first link share “earthquake”, the nodes of the second link share “history”, and the nodes of the last link share “hazard”. The *similarity* will be defined later.
- *The distance* between the subject heading and the entry in Wikipedia is *smaller*. In other words, the smaller the number of hops from the entry to the subject heading is, the more appropriate the subject heading is. This assumption is based on the intuition that too many hops from the entry will lead to induction of irrelevant subject headings.

Now we define the score of each subject heading. Given an induced subject heading H_i , the score of H is defined as follows:

$$\text{score}(H_i) = \text{score}(\text{Cmatch}(H_i)) \quad (1)$$

where $\text{Cmatch}(H_i)$ is a Wikipedia category which *matches* H_i . In this paper, we regard a Wikipedia category which has exactly equal string to H_i as a *matched* category.

Next, the score of each Wikipedia category C_j is recursively defined as follows:

$$\text{score}(C_j) = \max_{C_k \in \text{Cchild}(C_j)} \{1 - \alpha(1 - \text{sim}(C_j, C_k))\} \cdot \text{score}(C_k) \quad (2)$$

where $\text{Cchild}(C_j)$ is the collection of categories which have C_j as their parent (broader) category, α is the propagation parameter ($0 < \alpha < 1$), and $\text{sim}(C_j, C_k)$ is the similarity between two categories (defined below). The propagation parameter α reflects *the distance*, because the bigger α is, and the bigger the number of hops, the smaller the score is.

Finally, given a Wikipedia entry E , the scores of categories which are directly assigned to E are defined as follows:

$$\text{score}(C_i) = 1 - \alpha(1 - \text{sim}(C_i, E)) \quad (3)$$

where $C_i \in \text{Cparent}(E)$, and $\text{Cparent}(E)$ is the collection of categories which are assigned to E .

The similarity between two categories (or a category and the entry) is defined as follows:

$$\text{sim}(N_1, N_2) = \frac{m}{\sqrt{n_1} \sqrt{n_2}} \quad (4)$$

where N_1 and N_2 are the two categories (or the category and the entry), m is the number of 2-grams which are shared by both N_1 and N_2 , and n_p are the number of 2-grams which are contained in N_p . For example, given $N_1 = \text{“Dogs”}$ and $N_2 = \text{“Dog”}$,

- $n_1 = 3$ (“Do”, “og” and “gs”),
- $n_2 = 2$ (“Do” and “og”), and
- $m = 2$ (“Do” and “og”),

so the similarity is calculated as

$$\text{sim}(N_1, N_2) = \frac{2}{\sqrt{3} \sqrt{2}} \cong 0.8165$$

4.3 Algorithm

Based on the definition described in Subsection 4.2, subject headings which are related to a given keyword can be induced with scores. Our method induces subject headings from an inputted keyword as follows.

1. Retrieve a Wikipedia entry E of which title is most similar to the inputted keyword, based on the 2-gram model. The similarity is calculated using (4).
2. Give scores to categories which are directly assigned to E , using (3).
3. Propagate scores for broader categories which are assigned to the categories that have already given scores, using (2). Note that if the number of hops from E is unlimited, the propagation process will explode, because the numbers of broader categories for each node are usually multiple. To prevent the process from explosion, the maximum number

of hops from E is limited (currently we set the maximum number of hops to 5).

4. Retrieve subject headings which match exactly with the categories that are given scores. The scores of each retrieved subject heading are given using (1).

5. EXPERIMENTS

In order to estimate how our method adds values to subject headings, we applied it to the English Wikipedia and LCSH. The English Wikipedia has approximately 2.6 million articles, which is the largest one among various languages. LCSH is the de facto universal controlled vocabulary which has been adapted as a model for developing subject headings systems by many countries around the world.

5.1 Data

The entire data of the English Wikipedia was downloaded from the website of Wikimedia Foundation¹. We used the XML-formatted file of the English Wikipedia that was dumped in October 2008². This version has 2,605,674 articles.

The electrical version of LCSH was downloaded from LCSH.INFO³ on November 2008. The website provided linked-data of LCSH for experiment. The downloaded data contains 266,866 subject headings.

5.2 Coverage

If our method is considered as expansion of LCSH, any Wikipedia entry from which more than one subject heading are induced can be regarded as a *potential subject heading*. In other words, potential subject headings are associated with the LCSH system, so those are (at least potentially) useful for finding physical materials and classification (LCC). Therefore the number of the Wikipedia entries associated with LCSH will be a barometer of the added value to LCSH.

We applied the algorithm described in Subsection 4.3 to all the entries (the titles of 2,605,674 articles) of the English Wikipedia. As a result, 2,143,263 entries (82.3%) were associated with at least one subject heading in LCSH. The result shows that our method can potentially expand LCSH to nine times as large as the original size of LCSH.

5.3 Induced Subject Headings

Needless to say, all subject headings connected to the potential subject headings are not necessarily relevant. Usually multiple subject headings are induced for each potential subject heading, ranked with the calculated scores. The evaluation of induced subject headings is difficult, because there is no clear definition of *relevant* subject headings. The *relevance* depends on users, purposes, situations, and so on. Therefore we are planning to evaluate the relevance in real situations of information retrieval (see Section 6). In this section, we give some suggestions of our method, by showing some statistics and examples.

For each entry that was associated with at least one subject heading, our method induced 12.21 subject headings on average. Figure 3 and Figure 4 show the distribution of the numbers of induced subject headings. The result indicates that the half of the Wikipedia entries is associated with more than ten subject headings, suggesting that various aspects are reflected in the Wikipedia category system.

Table 1 shows some examples obtained by our method. For example, 84 subject headings such as “Suicide”, “Islam”, “Violence”, “Accident”, and “Terrorism” are induced from the entry “September 11 attacks”. It seems that these subject headings reflect the various aspects of the terrorist attacks. The second example “Subprime mortgage crisis” suggests the important advantage of Wikipedia, that is, speed of modification. The entry “Subprime mortgage crisis” was firstly added to the English Wikipedia in May 2007, and via the modification process of more than 3,000 edits, the entry was adequately associated with “Financial crises” and “Macroeconomics”. Note that LCSH has a subject heading “Subprime mortgage loans”, however, the subject heading has never been associated with “Financial crises”.

5.4 Theme Graph

Although most of the induced subject headings in Table 1 seem to be useful for information navigation, some subject headings are not easy to be interpreted. For example, it is difficult to know why the sixth subject heading “Religion” is induced from the entry “Israeli–Palestinian conflict”, without a lot of background knowledge about Israel and Palestinian issue. In order to find out why a subject heading is induced, the propagation paths from the entry to the subject heading may be useful.

To visualize the propagation paths, we implemented a module which outputs so-called a “theme graph” for each entry. The module uses an open source graph visualization software Graphviz⁴. Figure 5 shows the theme graph for the entry “Israeli–Palestinian conflict”. Using the graph, it is easy to know the reason why “Religion” is induced, because the graph shows that the Wikipedia categories “Arab-Israeli conflict”, “Islamic history”, and “History of religion” associate the entry with “Religion”. The graph also shows various aspects of the conflict, including race issues, political geography, and Judaism.

6. AN APPLICATION: LITTEL NAVIGATOR

To evaluate the effectiveness of our method, application to real situations of information retrieval is inevitable. We have been paying attentions for reference services of libraries, because reference services are 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 navigation system Littel Navigator, and operate the system in several university libraries in Japan.

Figure 6 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

¹ <http://download.wikimedia.org/>

² <http://download.wikimedia.org/enwiki/20081008/enwiki-20081008-pages-articles.xml.bz2>

³ <http://lcs.info/>

⁴ <http://www.graphviz.org/>

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.

7. DISCUSSION AND CONCLUSION

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

- **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 7 shows the usefulness of our approach.

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: top-down approaches and bottom-up approaches. Natural language processing techniques, including similarity calculation and acquisition synonyms, will contribute to enhancement of the potentials.

8. REFERENCES

- [1] Cohen, L. B. and Still J. 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.
- [2] Calhoun, K. 2006. The changing nature of the catalog and its integration with other discovery tools. Final report. 52p. retrieved from: <http://www.loc.gov/catdir/calhoun-report-final.pdf>
- [3] Voss, J. 2006. Collaborative thesaurus tagging the Wikipedia way. Collaborative Web Tagging Workshop. Retrieved from: <http://arxiv.org/abs/cs/0604036> (Last revised 27 Apr 2006, version v2)
- [4] Antelman, K. and Lynema, E. and Pace, A. K. 2006. Toward a 21st Century Library Catalog. *Information Technology and Libraries*, Vol. 25, No. 3, pp. 128-139. Retrieved from: <http://eprints.rclis.org/7332/>
- [5] Sadeh, T. 2007. Time for a change: new approaches for a new generation of library users. *New Library World*, Vol. 108, No. 7/8, pp. 307-316. DOI=10.1108/03074800710763608
- [6] OCLC Online Computer Library Center. 2005. Perceptions of Libraries and Information Resources: a Report to the OCLC Membership. Available at: <http://www.oclc.org/reports/2005perceptions.htm>
- [7] OCLC Online Compute Library Center. 2006. College Students' Perceptions of Libraries and Information Resources: a Report to the OCLC Membership. Available at: <http://www.oclc.org/reports/perceptionscollege.htm>
- [8] Lois Mai Chan. 2000. Exploiting LCSH, LCC, and DDC To Retrieve Networked Resources: Issues and Challenges. Retrieved from: http://www.loc.gov/catdir/bibcontrol/chan_paper.html

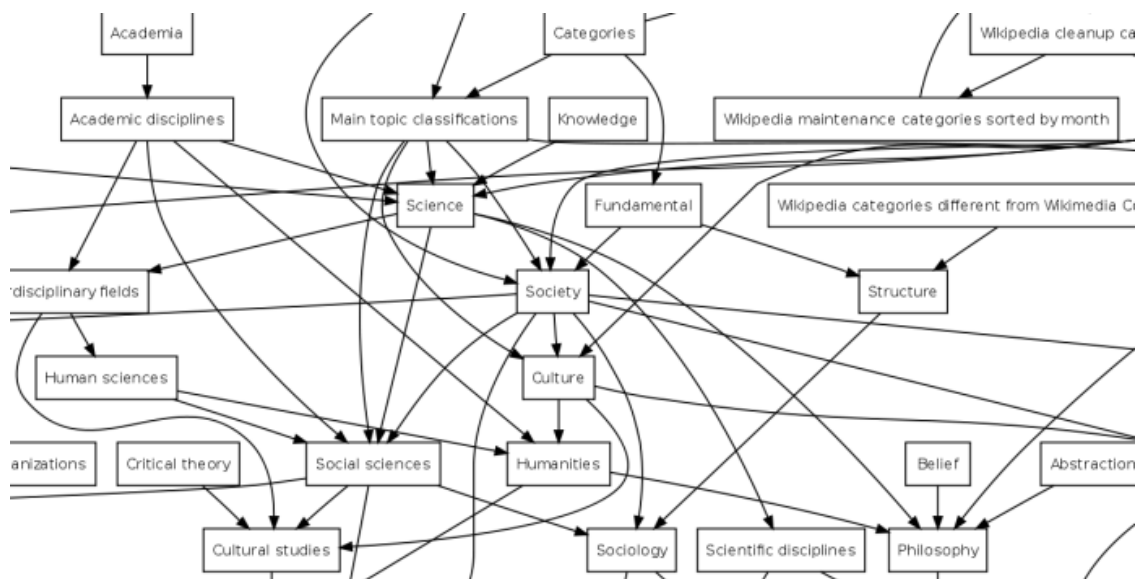


Figure 1: A snapshot of the structure of Wikipedia categories.

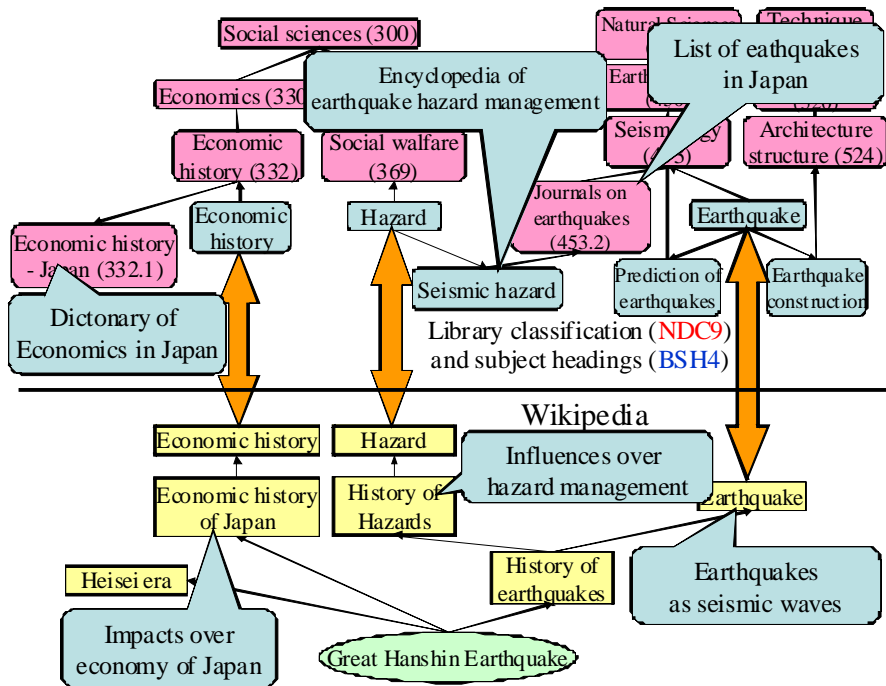


Figure 2: Induction of subject headings via the networks of Wikipedia categories.

Table 1: Examples of Induced Subject Headings of LCSH.

entry	top 8 subject headings	LCC	scores
September 11 attacks (84 subject headings)	Suicide	HV6543	0.1802
	Islam	BP1	0.1750
	Violence	HM886	0.1547
	Accidents	HB1323.A2	0.1521
	Death	BD443.8	0.1382
	Terrorism	HV6430	0.1349
	Massacres	BR1600	0.1349
	Transportation	GT5220	0.1324
Subprime mortgage crisis (34 subject headings)	Financial crises	HB3722	0.3722
	Economic history	HC	0.2267
	Economics	HB1	0.1612
	Macroeconomics	HB172.5	0.0957
	History	D	0.0774
	Money	GN450.5	0.0744
	Finance	HB1	0.0718
	Historiography	D13	0.0537
Super Mario 64 (27 subject headings)	Video games	GV1469.3	0.1587
	Games	GN454.8	0.0920
	History	D	0.0659
	Technology	T	0.0616
	Youth	HQ793	0.0524
	Mass media	P87	0.0477
	Communication	P87	0.0434
	Engineering	TA	0.0374
RIKEN MDGRAPE-3 (26 subject headings)	Supercomputers	QA76.88	0.2000
	Computer systems	QA75.5	0.0646
	Computer architecture	QA76.9.A73	0.0592
	Communication	P87	0.0544
	Computer engineering	TK7885	0.0405
	Computer science	QA75.5	0.0337
	Information technology	HC79.I55	0.0320
	History	D	0.0311
Israeli-Palestinian conflict (14 subject headings)	Arab-Israeli conflict	DS119.7	0.6739
	Islam	BP1	0.1250
	History	D	0.1122
	Political geography	JC319	0.0909
	Human geography	GF	0.0594
	Religion	BL48	0.0565
	Republics	JC421	0.0524
	Chronology	CE	0.0425

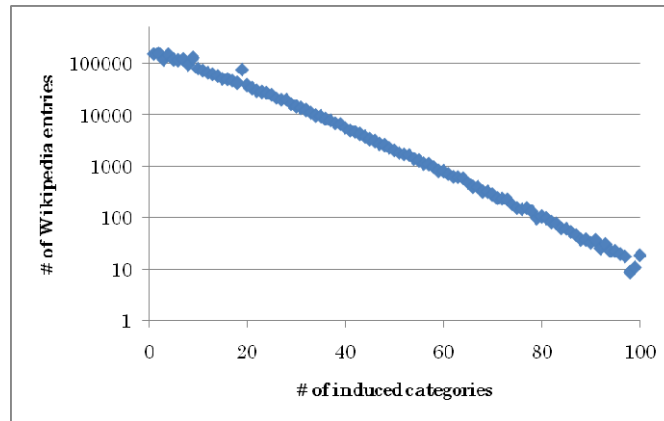


Figure 3: The distribution of Induced Subject Headings.

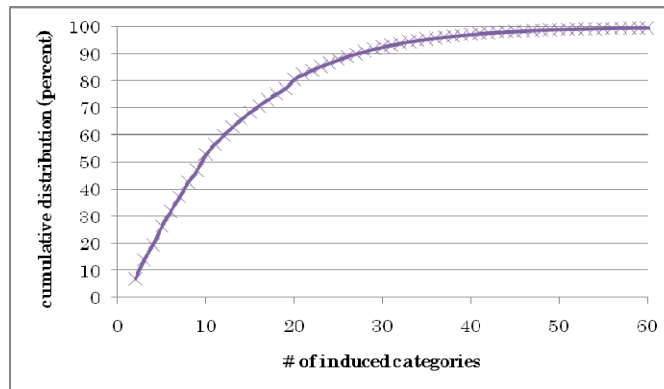


Figure 4: The Cumulative Distribution of Induced Subject Headings.

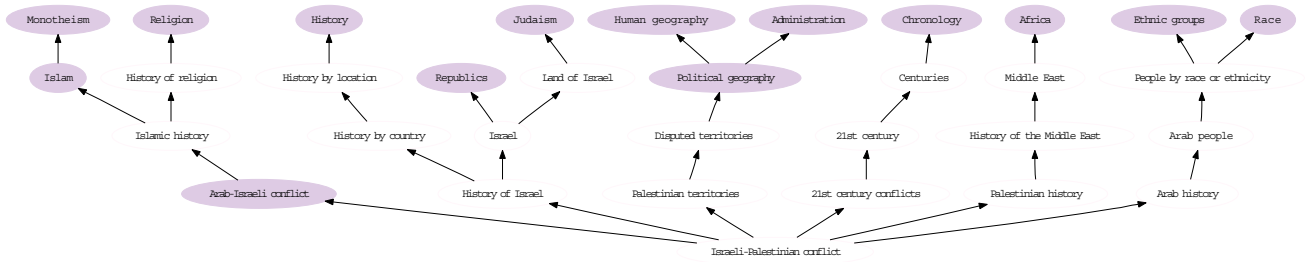


Figure 5: A Theme Graph.

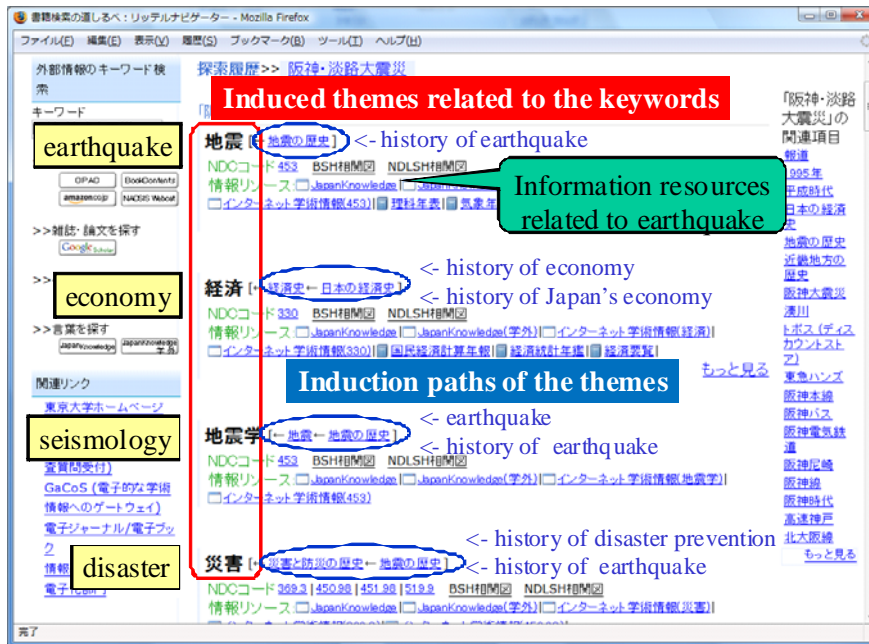


Figure 6: The screenshot of "Littel Navigator".

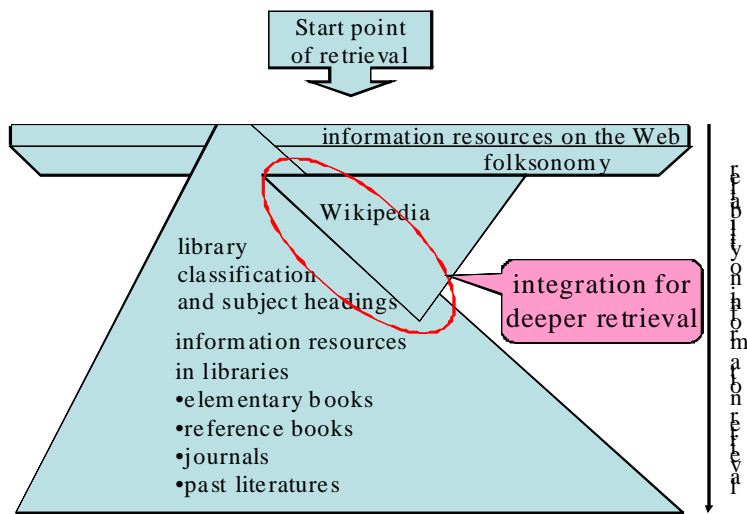


Figure 7: Navigation toward reliable information resources from any query keywords.