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

Yoji Kiyota, Hiroshi Nakagawa
Information Technology Center
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 Japan
kiyota@r.dl.itc.u-tokyo.ac.jp
n3@dl.itc.u-tokyo.ac.jp

Satoshi Sakai, Tatsuya Mori, Hidetaka Masuda
Graduate School of Science and Technology for Future Life, Tokyo Denki University
2-2 Kandanishikicho, Chiyoda-ku, Tokyo 101-8457 Japan
{sakai,mori}@cdl.im.dendai.ac.jp
masuda@im.dendai.ac.jp

ABSTRACT
This paper addresses an approach that integrates two different types of information resources: the Web and libraries. Our method begins from any keywords in Wikipedia, and induces related subject headings of LCSH through the Wikipedia category 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

1. INTRODUCTION
Although web search engines have several advantages on information retrieval, and have been used by a lot of ordinary people, 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. Several studies pointed out the importance of seamless integration of Web resources and library resources in order to activate advantages of both resources [1]. To realize such integration, the shape of classification systems (including LCSH, LCC, and so on) should be reexamined. Wikipedia categories, designed in order to organize enormous number of articles in Wikipedia, give us a clue for the integration. Voss [2] explored the category system of Wikipedia, and indicated that the system has a thesaurus-like structure that combines collaborative tagging (a bottom-up approach) and hierarchical subject indexing (a top-down approach) in a special way.

2. INDUCTION OF SUBJECT HEADINGS
Figure 1 shows the overview of our method. Suppose that we begin retrieval from a keyword “Hanshin Great Earthquake”. In the Japanese version of Wikipedia, the keyword 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 “Earthquakes”. As a result, we can get a subset of related categories as a tree structure. Wikipedia categories “Economic history”, “Hazard”, and “Earthquake” are correspondent with subject headings of BSH4 (Basic Subject Headings), which is developed by Japan Library Association.

We applied the method to the English version of Wikipedia and LCSH, and found that 82.3 % of articles (2.14 million articles) were associated with at least one subject heading in LCSH. Here are some examples:

[September 11 attacks]: “Suicide” (HV6543), “Islam” (BP1), “Violence” (HM886), “Accidents” (HB1323.A2), and “Terrorism” (HV6430)

[Subprime mortgage crisis]: “Financial crises” (HB3722), “Economic history” (HC), “Economics” (HB1), “Macroeconomics” (HB172.5), and “Money” (GN450.5)

JCDL'09, June 15-19, 2009, Austin, Texas, USA.
ACM 978-1-60558-322-8/09/06.

3. ACKNOWLEDGMENTS
This work was supported by KAKENHI Grant-in-Aid for Young Scientists (B), 20700128, 2008.

4. REFERENCES
[1] 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

[2] 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)

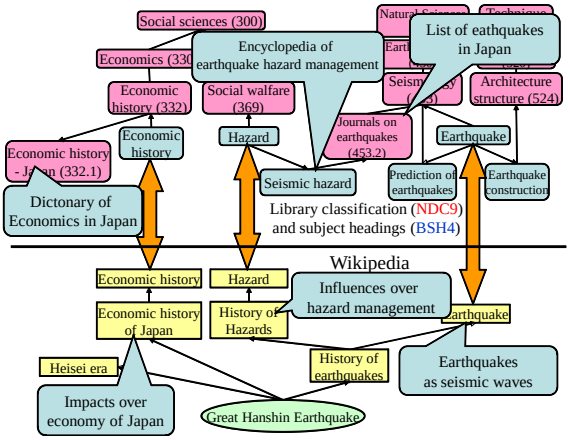

Figure 1: Induction of subject heading via Wikipedia categories.