Web Directories for Information Organization on Web Portals

Xin Fu
The University of North Carolina at Chapel Hill, USA

INTRODUCTION

Two methods are currently used to organize and retrieve information on the Internet. Search engines like Google and AltaVista use a robot-based keyword searching method by constructing inverted index files for Web pages and matching users' query terms with the index terms. The other method organizes human-selected Internet resources into a searchable database, and gives users structured hierarchical access to the database in a similar way to browsing through library classification schemes. We call this structured hierarchical system a Web directory. Knowledge structures, like a library classification schema or a Web directory, visualize and reflect what people know about things, and help people understand things better, identify gaps, recognize patterns, predict future trends, and so forth (Kwaśnik, 2005). Moreover, Web directories offer quality control and give access only to selected Internet resources. All these advantages make the browsing structure based on subject classification a desirable complement to the search engine type service (Koch, Day, Brümmer, Hiom, Peereboom, Poulter, & Worsfold, 1997).

Since the first widely known Web directory was constructed by Yahoo! in 1993, many such directories have been built up. Even the most popular robot-based search engines, such as Google and AltaVista, are also maintaining their own directories. On the other hand, many researchers have been trying to use traditional library classification schemes, such as Dewey Decimal Classification, to organize Internet resources. In the Dewey Decimal Classification (DDC) Online Project, Markey demonstrated the first implementation of a library classification scheme for end-user subject access, browsing, and display (Vizine-Goetz, 1999). Currently, not only the international general classification schemes (also called universal classification schemes), such as DDC, Universal Decimal Classification (UDC) and Library of Congress Classification (LCC), are employed[1], but also some national classification schemes[2] and subject-specific classification schemes[3]. Koch, Day, Brümmer et al. (1997) presented perhaps the most comprehensive study and comparison so far on the use of library classification schemes in organizing Internet information resources. They investigated three types of schemes, universal classification schemes, the national general schemes, and subject specific schemes, in terms of extent of usage, multilingual capability, strengths and weaknesses, integration between classification scheme and other systems (e.g. controlled subject headings), linking to third-party classification data, digital availability, copyright, and extensibility.

As Marcella and Newton noted, "the whole object of classification … is to create and preserve a subject order of maximum helpfulness to information seekers" (Van der Walt, 1998). At a time when both Internet-based classification schemes and traditional library classification systems are being used to provide access to Web resources, it is natural to compare the two and consider whether homegrown Web directories outperform the traditional library classification schemes in organizing information resources on the Internet. This will enable us to take advantage of their respective strengths and design more effective Web portals.

BACKGROUND

The literature about library classification exists in a huge volume. However, only a limited number of articles have addressed the topic of applying library classification schemes to organizing the information on the Internet. Likewise, not many authors have written about Web directories (compared to the vast pool of literature on automatic retrieval systems such as search engines). Even fewer have tried to juxtapose the two.

Among these trials, three articles are most related to the topic of this article. Van der Walt (1998) investigated some of the main structural features of the classification schemes used in the directories of search engines in order to determine whether they conform to the principles of library classification. The author examined 10 search engines at the main class level, analyzed the full hierarchies of a sample of three specific subjects in four of search engines, and identified a number of differences in the principles of constructing library classification scheme and Internet classifications. Ma (2001) compared the principles of designing traditional classification schemes and Web directories and pointed out some characteristics of the structure of Web directories. He noted that all the characteristics were determined by the Internet environment in which the directories functioned. Vizine-Goetz (1999) reviewed the major characteristics of DDC and LCC and assessed whether the electronic versions of these schemes could be successfully extended to the Internet. Through comparing Yahoo! and DDC classification, the

author concluded with some recommendations for improvements that online library classification schemes will need to make if they are to be used in the Internet environment.

Besides, some authors wrote about the influence of Colon Classification on Web directories. Chen and Fan (1999) analyzed the classification system used in the Yahoo directory and noted that it had a close relationship with the Colon Classification idea proposed by famous Indian classification scientist Ranganathan. Chan (2000) quoted Aimee Glassel to analyze the application of Colon Classification to Yahoo! and noted that "both systems are based on combining facets to facilitate searching and maximize the number of relevant results." It was argued that "Ranganathan's ideas of classification are more applicable now than before in the Internet environment."

In this article, the author will study the structure of current Web directories and compare it with major universal library classifications. Focus will be on their main classes with some additional discussions on hierarchical structures. The study does not emphasize a specific Web directory or a specific library classification scheme; instead, it refers to a number of Web directories and library classification schemes as examples to support the arguments. Considering the scope of this article, only the comprehensive Web directories used in major Web portals and the universal classification schemes (like DDC, LCC and UDC) will be studied.

WEB DIRECTORIES VS. LIBRARY CLASSIFICATIONS

Comparison

Web Directories

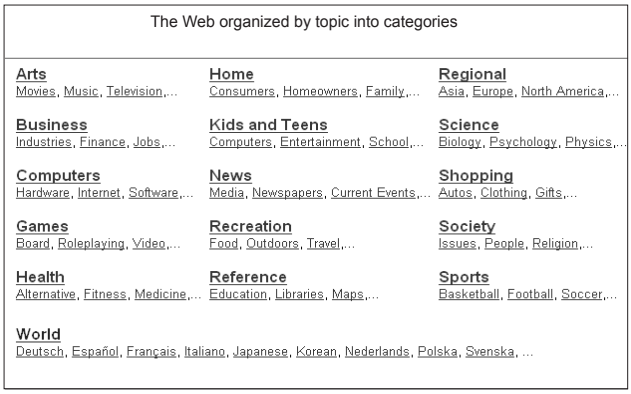
Figures 1 and 2 display the first page of Yahoo! directory and Google directory, and Table 1 is a mapping between them. It can be easily noted that the two classification schemes match quite well at the main class level. All the Yahoo! main classes except "Government" can find their counterparts in Google main classes. Conversely, all the Google main classes except "Home" and "Kids and Teens" have their counterparts at Yahoo!'s main classes. As a matter of fact, the main classes in most other Web directories are organized in a similar way, so the differences between Web directories can be neglected when comparison is made with traditional library classifications.

Library Classification Schemes

Figure 3 and Figure 4 display the main classes of Dewey Decimal Classification scheme and Library of Congress Classification scheme. Table 2 compares the two. Again, the main classes in these two classification schemes match quite well with each other. Differences between them will also be neglected when they are compared with Web directories.

Figure 1. The main classes in Yahoo! Directory (Retrieved December 8, 2003, from http://www.yahoo.com)

Figure 2. The main classes in Google Directory (Retrieved December 8, 2003, from http://www.google.com/dirhp?hl=en&tab=wd&ie=UTF-8&oe=UTF-8&q=)



What is Different?

Unlike the high degree of consistency between Web directory main classes and between library classification main classes, a comparison between the Yahoo! classification and the LCC scheme reveals tremendous differences at main class level.

The first, and the most obvious difference is that only four of the Yahoo! main classes coincide with main classes in the UDC scheme or LCC scheme: "*Arts & Humanities,*" "*Science,*" and "*Social Science*" in Yahoo! with 700 "The