Two Digital Library Interfaces That Exploit Hierarchical Structure

Robert B. Allen
Bellcore
MRE 2A367
445 South Street
Morristown, NJ USA
rba@bellcore.com

1. ABSTRACT

Two library classification system interfaces have been implemented for navigating and searching large collections of document and book records. One interface allows the user to browse book records organized by the Dewey Decimal Classification hierarchy. A Book Shelf display reflects the facet position in the classification hierarchy during browsing, and it dynamically updates to reflect search hits and attribute selections. The other interface provides access to records describing computer science documents classified by the *ACM Computing Reviews* (CR) system. The CR classification system is a type of faceted classification in which documents can appear at several points in the hierarchy. These two interfaces demonstrate that classification structure can be effectively utilized for organizing digital libraries and, potentially, collections of Internet-wide information services.

2. CLASSIFICATION SYSTEMS FOR ORGANIZING LARGE ELECTRONIC INFORMATION ARCHIVES

2.1. Advantages of Classification-Based Interfaces

Organizing books and documents in a digital library interface by an a priori classification system may seem to be a weak alternative to the variety of ad hoc organizations possible in response to searches. However, a consistent structure, reflecting a commonly agreed upon organization of knowledge, may help orient the user. As suggested by Mann[15]:

> *Given identical computer systems for searching the catalog records, is there an additional and substantial advantage in being able to search the full texts themselves in subject-browseable groups?*
>
> *I submit that anyone who actually has to do research, especially in unfamiliar subject areas or in languages in which he* [sic] *has little proficiency, would have a decided and fully justified preference for working in Library A* [with subject-browseable groups]. (page 131).

Indeed, an interface which reflects the structure of the classification system essentially provides suggestions to a user about further options to pursue following a search. That is, after a search a user can select from the node labels of the classification system near the search hits to identify the subdivisions that may help further refine the search. The classification system can also be used to restrict searches so as to reduce the computational cost and avoid overwhelming users with spurious information.

This paper considers two types of interfaces for accessing books and documents organized in classification systems. The interfaces have been implemented in the X Window System using Motif widgets. The first interface (Section 2) is for the Dewey Decimal Classification (DDC). this uses the hierarchical organization to facilitate browsing and the presentation of book records. The second interface (Section 3) manages documents organized by a type of faceted classification system.

2.2. OPACs and Electronic Book Interfaces

Several interfaces have been developed for accessing online book records. However, most Online Public Access Catalog (OPAC) interfaces are designed for ASCII terminals and do not have advantages, such as direct manipulation, associated with GUIs. Other OPACs provide extensive term searching but do not take advantage of the hierarchical organization [8]. Book cataloging systems also provide access to the hierarchical classifications. However, these generally have only simple graphical interfaces (e.g., [18]) and are not documented in the literature. Some prototype electronic catalogs introduce creative interfaces but may not scale well for large collections [4, 17, 19].

Interfaces for electronic books have now been widely studied, but relatively little attention has been paid to the management of collections of books in these systems. The SuperBook^{TM} browser [6, 7] takes advantage of the hierarchical structure of individual documents. For instance, it presents chapter and section headings in a dynamic Table of Contents (TOC). However, the SuperBook browser itself is not effective for navigating a hierarchical book classification system; it does not easily support fielded search, and it is not designed for presenting and manipulating short records.

Section 2 describes an interface that incorporates interface features from other systems and adds many new ones. Among these features are fisheye browsing of the

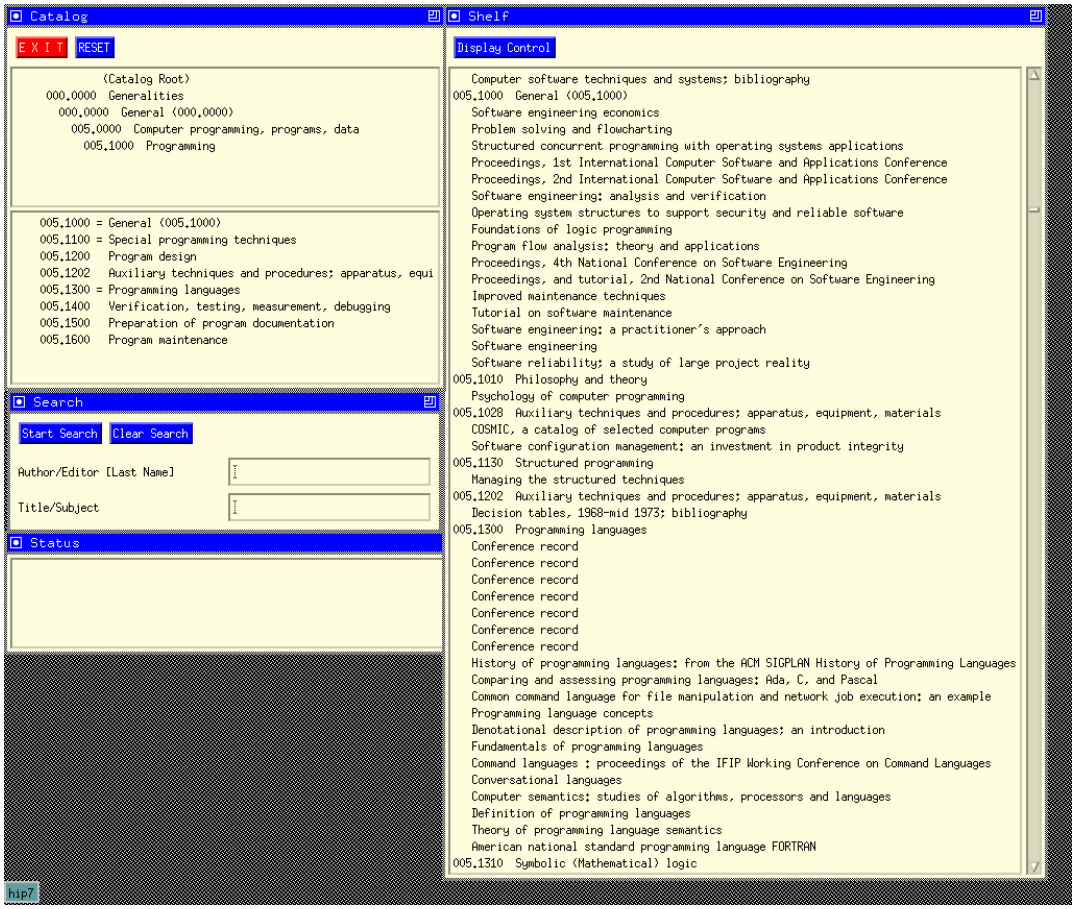

Figure 1: GUI for Book Records Organized by Dewey Decimal Classification.

classification hierarchy, a full Book Shelf, interlocking operation of the classification hierarchy and Book Shelf display, posting search hits against the classification hierarchy, control of search hit displays on the Shelf, control of the granularity of the search hit displays, and lateral links across the classification hierarchy. Moreover, it supports a realistically large collection of book records.

2.3. Interface for Faceted Classifications

Many classification hierarchies have multiple components. These include faceted classifications [21], polyhierarchies, and multitrees [10]. Faceted classifications are the most widely explored of these systems and they have been proposed as suitable for online retrieval by Godert [11]; but electronic systems to manage these have not been previously described. Because documents may be included under several different nodes of a faceted classifications, the faceted classifications are a type of directed acyclic graph. On the other hand, any faceted classification can be expanded as a simple hierarchy.

Some classification systems are partially faceted. For instance, books in the DDC under Art History are organized by geographic areas and historical periods. Books organized by the Library of Congress system include Cutter number extensions which are orthogonal to the main classifications. Many other classification systems, such as the INSPEC Classification for engineering and the *ACM Computing Reviews* (CR) classification system [2] are faceted.

3. HIERARCHICAL-CLASSIFICATION INTERFACE

Figure 1 shows an interface that allows interaction with the DDC. The user has navigated the Subject Hierarchy List to **005.1000 Programming**. The interface is composed of three main groups of widgets which are described below.

3.1. Interface Widgets

3.1.1. Book Records and the Dewey Decimal Classification The DDC probably is the most widely used international classification system. It is also one of the purest hierarchies of the major library classification systems. The DDC was designed for cataloging books [18], but it has been suggested as the basis for an interface to help the casual user [16]. With the introduction of high-powered personal workstations and flexible GUIs, the accomplishment of this goal for the casual user is now feasible. The headings for a large part of the DDC were obtained and merged with the book records. While the DDC, as with any classification system, is not suitable for all tasks, it is useful for a large range of tasks and is familiar to many users. In preparing the corpus,

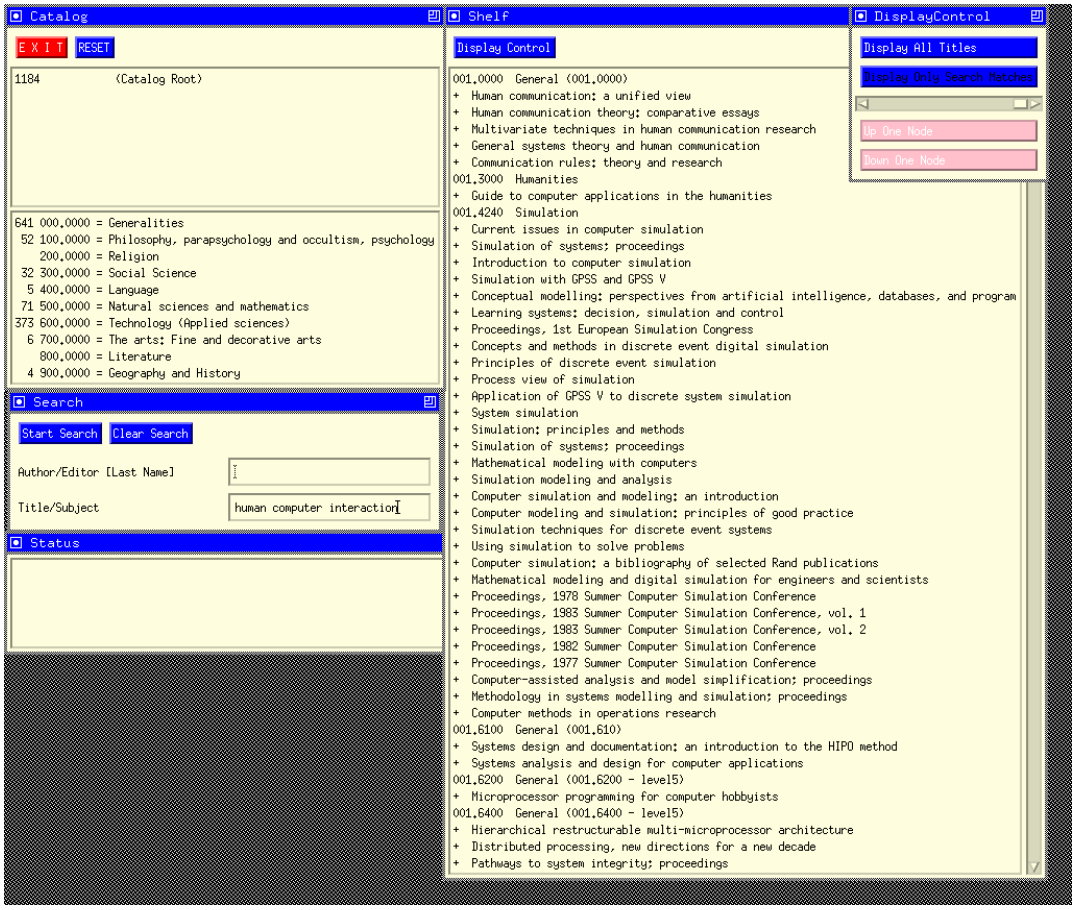

Figure 2: Interface after Search for “Human Computer Interaction”.

long call numbers were truncated to 4 decimal places. In a few cases, the hierarchy was not complete and filler headings were inserted. For instance, in the Classification immediately below the first-level node **000.0 Generalities** is the third-level node **001.0 Knowledge**. A second-level heading **000.0 General** was created to match other second-level headings under **000.0 Generalities** such as **010.0 Bibliography**.

Book and document records numbered by the DDC were obtained from the Bellcore Technical Libraries. They covered approximately 50,000 books and technical reports. Each record included the shelf number, author, title, publisher, location, a subject field, and a list of the library locations where the book was held.

3.1.2. Subject Hierarchy and Current Node Lists: The upper left quadrant of Figures 1 and 2 shows a TOC for the hierarchical interface. The TOC is split across Subject Hierarchy and Current Node Lists. Together, these widgets allow a user to navigate through the hierarchy and serve a function similar to the expandable TOC of the SuperBook browser. In a deep and wide hierarchy, such as the DDC, the contents of the expanding TOC would frequently scroll out of view. Although less information is presented in separate Subject Hierarchy and Current Node Lists than in an expanding TOC, these lists yield a more predictable display and are especially suitable for the DDC records where the shelf number provides an additional pointer into the hierarchy. Moreover, book-record hierarchies have looser semantic connections between nodes at the same level than the TOCs of most individual documents and books. Thus, displaying all choices at intermediate-level nodes would not be particularly informative.

3.1.3. Book Shelf and Book Display Widgets: The Book Shelf (right side in Figures 1 and 2) does not attempt to mimic a physical book shelf. Rather, it is a very long list of book records. The user, typically, has only a partial view of the list. The view of the Shelf is limited by the number of items that can be displayed on the screen at any time and by options that determine which book records and which attributes of those records are to be displayed.

The selection of displayed attributes is determined in response to iterative queries that control a filter mask. Thus, the Book Shelf is “dynamic” in the same sense as the dynamic graphical query interface described in [22] and as used in general purpose data viewers (e.g., [20]). Nodes in the classification system immediately above the selected books are also presented on the Shelf. The Shelf shows nodes at different levels abutted one after

the other. The default display for records on the Shelf shows titles. The user can select other attributes to be presented on the Shelf such as the author name, the length (number of paper pages), and the publisher.

When the user clicks on a book title on the Shelf, a Book Display widget opens showing the full record for that book. Indeed, it is possible to browse the Shelf by selecting successive book titles to be displayed.

3.1.4. Fielded Search Widget: The Fielded Search widget (lower left in Figure 1) generates searches on book record fields such as title, author, and subject descriptors. Three search algorithms are available: a Boolean OR of matched terms, term matches between the query and the document terms weighted by term frequencies, and Latent Semantic Indexing (LSI) [5].

For LSI [5] searches, the LSI-value for a node is derived from the position of all the terms in the book titles and subject descriptions of all the books under that node. This is conceptually similar to the approaches of [9, 13] for other search algorithms. However, it meant that individual books were not able to be located with LSI. Moreover, because the LSI searches took considerable computational resources for matching vectors, the LSI space had to be precomputed.

3.2. Browsing

The interface can be used for browsing the DDC. The Current Node List displays items that allow the user to navigate deeper into the hierarchy. Initially, the current nodes are the top-level classification terms (as shown in Figure 1). When nodes lower in the hierarchy exist, the nodes above this are marked with an "=". The Subject Hierarchy List displays the hierarchy nodes above the books currently being displayed on the Book Shelf. Clicking on one of the higher-level nodes causes the immediate descendants of the selected node to be displayed in the Current Node List. In addition, the Shelf displays books at the selected node.

3.3. Searching

Figure 2 shows the interface following a search on the terms "Human Computer Interaction". Titles that match the search are marked with a "+". In the default Hits Only display mode, the Shelf displays only the matched books and their immediate parent nodes. However, the DisplayAllTitles button (at the upper right) lets the user display all titles with the hits interspersed.

Counts of search matches are posted beside the node labels on the TOC widgets. These counts can help the user locate relevant items. For instance, in Figure 2 1184 books match the query and 641 of these are under the heading **000.0 Generalities**. This suggests that is the most promising part of the hierarchy for looking for relevant books.

The hierarchical interface is most effective for comparing documents of relatively similar retrieval values because it does not easily display quantitative information about the matches. That is, unlike typical information retrieval (IR) systems that present items ranked by a similarity metric, the interface based on hierarchical structure does not readily show graded retrieval scores. The approach taken here is to set a threshold in the ranked-ordered list and to treat all items above that threshold as hits. Initially, a titration procedure was developed to select the threshold so that, not less than 5 titles and not more than 100 titles would be presented. However, informal user testing suggested that users often wanted to override the titration setting. Thus, a slider for controlling the number of hits displayed was developed. This is similar to the use of a slider for "aggregation manipulation" [12]. In Figure 2, the slider (upper right) has been positioned to show the maximum number of hits (1184 in this example).

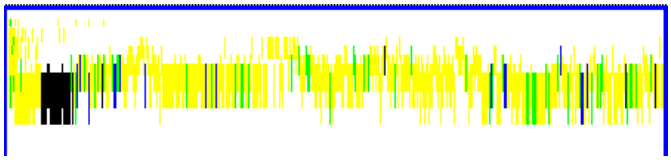

Figure 3: Graphic Display of Dewey Hierarchy after LSI Search.

It has been believed that book titles are too short to yield effective searches. However, the assumption behind this work is that there are often enough records in a node that relevant words will appear in, at least, some of them. Getting search matches on some of the titles in a node allows the user to reach that node and then to use the Shelf browsing capability of the interface and then to find the most relevant documents. In addition, following a search, the user could easily step forward and backward on the Shelf with the NextMatchNode and PreviousMatchNode buttons.

3.4. Extended Features

Several additional features were implemented for the hierarchical interface but were not included in the basic version.

3.4.1. Interactive Graphic view of DDC: Graphics can often help orient users with large amounts of data. However, graphical displays have been only lightly used in information interfaces [14]. Figure 3 shows a black-and-white view of a compressed dendrogram of the nodes in the Dewey hierarchy. Like [3], this dendrogram is interactive. In this case, clicking on the dendrogram causes the Book Shelf, Subject Hierarchy and Current Node Lists to open to the selected node. The dendrogram in Figure 3 shows search hits from an LSI search on the term "computer". Dark lines indicate better matches. Clearly, many of the computer-related books are in the early part of the hierarchy. The graphic display tool is still in early stages of development. For instance, the node representations are so closely spaced that it is difficult to see them and to select them.

3.4.2. Restricting Shelf by Attributes: Attributes, such as library location, whether the document has been checked

out, and the type of document, may be used to select subsets of books controlled by menus. By selecting various library locations it is possible to examine the virtual Shelf for any one location or any combination of locations of the Bellcore Technical Libraries.

3.4.3. Additional Shelf Traversal Modes: Two additional modes for skipping through search hits on the Book Shelf were implemented. It was possible to skip by Book and by search-algorithm-match order. Specifically, the UpBook and DownBook buttons allow the user to easily find book titles that match a search. The PreviousBookInOrder and NextBookInOrder buttons let the user examine books in the ranked order in which they matched the query. However, it is easy for the user to lose orientation because the books are not necessarily in order and the user viewing them may jump around the hierarchy. If the user requests NextBookInOrder after all the books in the initial set have been viewed, the set expands by relaxing the threshold.

3.4.4. Similar Books: An option for the Book Display allows the user to request Similar Books. It searches for books similar to the displayed book where similarity is determined by one of the retrieval algorithms rather than by shelf proximity. This option spawns a new search that, when it follows an initial search, is a type of relevance feedback. Because the book records are short, the Similar Book requests yield some spurious matches. As with the initial searches, posting similar-book hits against the Subject Hierarchy List allows the user to follow the classification semantics to identify relevant items. The Book Display also contains options for presenting other books by the same author. This links books across leaf nodes of the hierarchy.

3.4.5. Lateral Links: For especially complex hierarchies, when a person using the browser reaches a terminal node they may not find exactly the information they are looking for but they may suspect they are close to it. Requesting a search for Similar Books (see above) would be one way to find other relevant sections of the hierarchy, but it is also possible to have precomputed lateral links between nodes (i.e., "distributed relatives"). A mechanism was implemented for this, in which a button was associated with each node and clicking on that button presented a list of other related nodes. At some point, these complex hierarchies would be better represented by faceted classification systems (see Section 3).

3.4.6. User Restricted Collections: In many cases, a user would be willing to restrict searches to certain segments of the classification hierarchy. This could improve computational efficiency and would focus the users attention. While that capability was not essential for the current prototype with about 50K book records on a powerful workstation, for a much larger collection (e.g., for the Library of Congress collection or for World-Wide Web (WWW) pages on the Internet) the user should be able to specify subsets of the records to search. For this system, users sub-selected nodes to include on a separate shelf and they could toggle back and forth to that shelf.

4. INTERFACE FOR FACETED CLASSIFICATIONS

Figure 3 shows an interface for browsing the computer science literature by means of the *Computing Reviews* classification. The test corpus consisted of doctoral dissertations cited in *ACM Computing Archive* [1] as published in 1992. The key idea is selection by specifying multiple constraints. Of course, there is no linear organization of documents for display in this collection; thus, the order of the nodes in the shelf displays is undetermined.

4.1. Interface Widgets

4.1.1. Cascading Facet Menus and Active Constraints Widget: Major categories are chosen from the Facets widget at the upper left of Figure 4. These selections open cascaded menus that display lower-level categories. When the "+" to the right of the facet label is selected, the facet is added to the Current Constraint List (left middle in Figure 4).

To show the context of the selected constraint labels, the parents of the constraints are displayed in parentheses on the Constraint List. The Shelf is updated with articles that match the constraints. Of course, the constraints propagate to all their descendants. Constraints can be dropped from the Constraint List by clicking on the "–" on the right side of the widget.

The interface allows the user either to take documents that match the union of the constraints (AND) or the intersection of the constraints (OR). For large collections, there are often far too many matches for the union. By switching to the AND display, the most relevant documents can be easily found. For the ACM CR collection, there is substantial variability in the number of categories assigned and the criteria for determining relevance of those categories.

4.1.2. Shelf: Because most of the documents are assigned to several categories, a user could find a relevant node and then find other nodes that have similar classifications. The overlapping categories are presented in the current interface by selecting the "o" from the first vector on the right side of the Facet Menu widget.

Among doctoral dissertations that were cited in *ACM Computing Archive* [1] as published in 1992, the categories that had two or more overlaps to **H.3.3 Information Storage and Retrieval** were **H.2.4 Systems**, **H.2.0 General**, **D.3.2 Design Styles**, **H.5.2 User Interfaces**, and **I.2.6 Learning**. Thus, a user who accessed articles under **H.3.3** could examine those other categories for relevant material. This is a type of *lateral link* across the hierarchy (see "Extended Features" section above).

4.1.3. Searches: Currently, term-frequency weighted searches are implemented in this interface. In one mode, it is possible to ask for all document titles to be included in the search. It is also possible to limit the search to those documents that match the constraints.

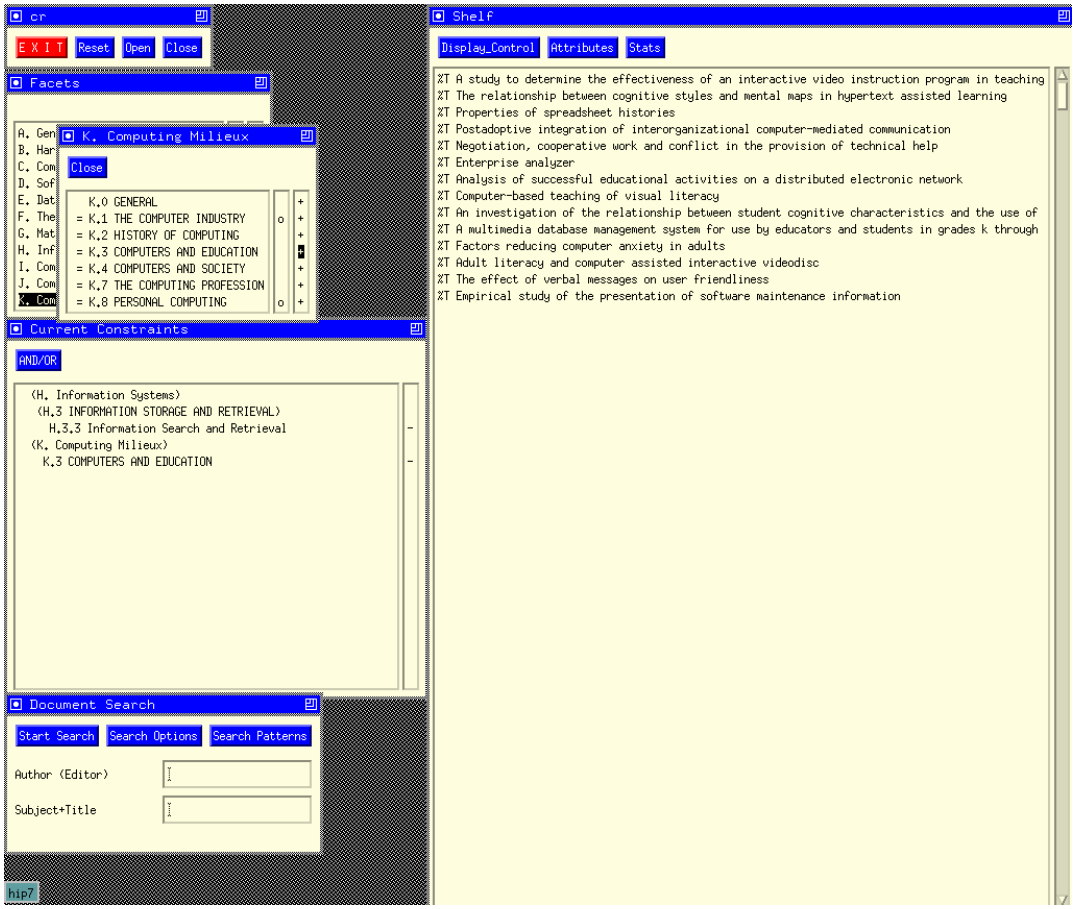

Figure 4: Interface for *Computing Reviews* Classification with Two Constraints Selected.

Posting search hits against the hierarchy is more complicated in this case than for the simple hierarchical display because a single document can belong to several categories. The current system uses fractional category memberships when the hits are spread across categories. As noted above, the Book Shelf for the facet interface has no a priori order. Thus, there is no natural order to display search hits. On the other hand, a variety of other ad hoc organizations are possible. For instance, the categories might be ordered by the density of hits. A related problem is which facet hierarchy to pop-open after a search (perhaps to help guide the user to further refine the search).

5. DISCUSSION

5.1. User Studies

While formal user studies have not been conducted on these interfaces, informal feedback from users of the hierarchical interface has been generally favorable. One major innovation here has been the introduction of a Book Shelf. Because this is the only full-scale system to include a Book Shelf (and hence the only system to allow browsing of books by shelf order), it is not clear what sort of evaluation is most reasonable.

The greatest problem with these interfaces appears to be complex interactions among features. For instance, in the Hits Only mode there are often too few selections to fill the Shelf Display; thus, the UpBook and DownBook buttons have no effect. In addition, some test users have suggested that the elision in the Hits Only mode should apply to the TOC as well as the Book Shelf. Completely shifting context from one set of screens to another (e.g., with the similar books option) is also difficult.

Beyond the problems of the interface design, there are limitations inherent in this type of interface for hierarchical classification systems. A substantial concern is the user does not know how many books are included under each node. For parts of the hierarchy hierarchy, a user may know or may be able to take a good guess; however, the user may not be at all familiar with other parts of the hierarchy.

The facet interface is probably harder to use than the simple hierarchical interface. This is because of the complexity of managing multiple facet hierarchies and the lack of a natural shelf order for the documents. Moreover, the facet interface described here has not been as well developed as the simple hierarchical interface. For instance, graphical displays might be especially useful for navigation of the facet hierarchies.

5.2. Integration with Other Information Systems

These interfaces could provide the basis for access to additional electronic information sources. Clearly, it would be possible to have the short document records pointing to the full text of the books and documents. Moreover, encyclopedia articles describing authors could easily be presented. Likewise, book reviews, citation statistics, circulation data, and user annotations could be included as part of the Book Display. Conversely, an electronic encyclopedia could access the OPAC for bibliographies.

Overall, these interfaces suggest that the structure of a classification system can be a useful aid for searching and navigating a digital library. Indeed, it may be worth exploring how digital library classifications can be extended to finding information in less structured domains such as for information in the WWW.

5.3. Envoi

Techniques such as the PreviousMatchNode/NextMatchNode buttons and lateral linking show how search-based IR and structure-based Hypertext approaches can be combined. It is also worth noting that structure could be used to enhance a search-based OPAC (e.g., [8]). In any event, while the DDC provides links to related documents, there are many other dimensions of similarity (e.g., author, citations, publisher) that could be used for linking as well. It remains to be seen whether these dimensions can be coordinated into useful interfaces.

ACKNOWLEDGMENTS

The DDC was used with the permission of the Online Computer Library Center (OCLC). The collection of book records used here was developed for test purposes and is not a Bellcore product. A much earlier version of this paper appeared in *Digital Libraries'94*, College Station, TX, June, 1994.

REFERENCES

1. ACM, *ACM Computing Archive*, 1994, New York.

2. ACM, ACM Computing Reviews Classification System. *ACM Computing Reviews 35* (1994) 4-44.

3. Allen, R.B., Obry, P., and Littman, M., An Interface for Navigating Clustered Document Sets Returned by Queries. *Proceedings of SIGOIS* (Milpitas, CA, June) ACM, New York, 1993, 203-208.

4. Borgman, C.L., Walter, V.A., Rosenberg, J.B., and Gallagher, A.L., Children's Use of a Direct Manipulation Library Catalog. *ACM SIGCHI Bulletin 23*, 4(Oct. 1991) 69-70.

5. Deerwester, S., Dumais, S., Furnas, G., Landauer, T.K., and Harshman, R., Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41* (1990), 391-407.

6. Egan, D., Lesk, M.E., Ketchum, D., Lochbaum, C.C., Remde, J.R., and Landauer, T.K., Hypertext for the Electronic Library? CORE Sample Results. *Hypertext '89* (Pittsburgh, Nov.) ACM, New York, 1989, 299-312.

7. Egan, D., Remde, J.R., Gomez, L.M., Landauer, T.K., Eberhardt, J., and Lochbaum, C.C., Formative Design and Evaluation of SuperBook. *ACM Transactions on Information Systems 7* (1989) 30-57.

8. Fox, E.A., France, R.K., Sahle, E., Daoud, A., and Cline, B.E., Development of a Modern OPAC: From REVTOLC to MARIAN. *Proceedings of SIGIR'93* (Pittsburgh, June) ACM, New York, 1993, 248-259.

9. Frisse, M.E., Cousins, S.B., and Hassan, S., WALT: A Research Environment for Medical Hypertext. *Hypertext'92* (San Antonio, Nov.) ACM, New York, 1992, 389-394.

10. Furnas, G.W. and Zacks, J., Multitrees: Enriching and Reusing Hierarchical Structure. *ACM SIGCHI'93* (Boston, Apr.), ACM, New York, 1993, 330-336.

11. Godert, W., Facet Classification in Online Retrieval. *International Classification 18* (1991) 98-109.

12. Goldstein, J. and Roth, S.F., Using Aggregation and Dynamic Queries for Exploring Large Data Sets. *ACM SIGCHI'93* (Boston, Apr.), ACM, New York, 1993, 23-29.

13. Hearst, M. and Plaunt, C., Subtopic Structuring for Full-length Document Access. *Proceedings SIGIR'93* (Pittsburgh, June), ACM, New York, 1993, 59-68.

14. Lesk, M.E., What To Do When There's Too Much Information? *Hypertext '89* (Pittsburgh, Nov.) ACM, New York, 1989, 305-318.

15. Mann, T., *Library Research Models*, New York, Oxford University Press, 1993.

16. Markey, K. and Demeyer, A.N., *Dewey Decimal Classification Online Project: Evaluation of Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog*, OCLC, Dublin OH, 1986, OPR/RR-86-1.

17. Micco, M. and Basista, T., Beyond Subject Access: The Next Generation of OPAC Software. *Proceedings Integrated Online Library Systems* (1991), 103-112.

18. OCLC (Forrest Press), *Electronic Dewey*. Dublin OH, 1993.

19. Pejtersen, A.M., A Library System for Information Retrieval Based on a Cognitive Task Analysis and Supported by an Icon-Based Interface. *Proceedings of SIGIR'89* (Cambridge, MA, June) ACM, New York, 1989, 40-47.

20. Swayne, D.F., Cook, D., and Buja, A., Interactive Dynamic Graphics in the Xwindow System with a Link to S. *Proceedings of the Section on Statistical Graphics of the American Statistical Association* (Atlanta) ASA, 1991, 1-8.

21. Vickery, B.C., *Faceted Classification.* New Brunswick, NJ, Rutgers University Press, 1965.

22. Williamson, C. and Shneiderman, B., The Dynamic HomeFinder: Evaluating Dynamic Queries in a Real-Estate Information Exploration System. *Proceedings of SIGIR'92* (Copenhagen, June) ACM, New York, 1992, 338-346.