

Creating an Electronic Databases Selection Expert System

Wei Ma

Asst. Reference Librarian & Asst. Professor of Library Administration

Timothy W. Cole

Interim Mathematics Librarian & Assoc. Professor of Library Administration

University of Illinois at Urbana-Champaign

Abstract

The past decade has seen an explosion in the numbers and types of electronic information resources available to identify and retrieve articles and other information relevant to a scholar's research needs. Today library users are confronted with a myriad of online and CD-ROM bibliographic and full-text databases from which they are expected to select the one or ones most germane to their information needs. The task for the academic library is to facilitate the selection process. That many users are now accessing Library-provided menus and database lists from outside the Library and at hours when no library staff member is available to assist, makes the challenge that much more difficult. This presentation will describe the creation and initial testing of an online, prototype, end-user database selection tool developed at the University of Illinois at Urbana-Champaign (UIUC). We will detail the environment, issues, and concerns that gave rise to our project, its relation to other work, and the approach we took and why. The prototype Web-based selection interface developed allows users to search for relevant databases by keywords and phrases taken directly from their search topics, by browsing librarian-assigned subject categories, and/or by identifying desired database characteristics. Behind this interface is an SQL database containing in-depth information about the characteristics of each database, including its complete controlled vocabulary (when available) or an extensive sampling of controlled vocabulary terms (or equivalent information) drawn from actual records.

Challenges and issues

When users need to find relevant information in today's computer-based information retrieval environment, they are first faced with the task of identifying the appropriate resource database(s) to search. This process has been complicated with the rapid development of information technology and the proliferation of digital information resources. At UIUC we've gone from having only a handful of single-workstation CD-ROM databases accessible to end-users in 1990 to having over 100 online journal article databases (accessible campus-wide) and over 200 CD-ROM databases available to end-users in 2000. More than half of the growth in available resources has taken place during the last 3 years.

The situation is exacerbated by constant change. Software migration, updates, product changes, vendor mergers, changes in database producer-vendor agreements, changes in database availability or cost, have all contributed to constantly changing search interfaces, database coverage, database features, and the selection of databases available at any time. It's difficult for Library faculty and staff to keep up; it's unrealistic to expect users to keep up with such constant change.

At UIUC, the highly distributed nature of the Library system (more than 40 separate public-service points distributed in more than 20 campus buildings) further complicates the problem. Most CD-ROM databases are available for searching at only 1 location, some at multiple libraries, and only a few are available at all library locations. This makes it all the more difficult for end users to know of the existence of all the electronic resources available system wide.

Nonetheless, demand for and use of these resources is high. In terms of number of searches performed, the combined end-user use of journal article index databases available campus-wide has exceeded total use of the online public access catalog by UIUC Library patrons for the past several years. User surveys done at UIUC confirm that these resources are highly valued. They also confirm that users are struggling to find the best resources to use for specific information needs. The result is that these resources aren't always being utilized to best advantage.

Current Practice

Library user observations suggest that patrons who used certain databases before have a tendency to select just those familiar databases, ignoring others which could be more appropriate for their topics (Meyer, 1990). New and remote users may feel confused when professional assistance is not readily available. Informal observations by library staff indicate that many users waste hours searching irrelevant databases. Even professional librarians with previous experience in selecting from among available electronic resources are having difficulty maintaining their expertise (Zahir, 1992; Thornburg, 1987). This further encourages heavy use of general-interest databases, and the overlooking of other resources (e.g., more specialized or in different locations) that might be more appropriate and relevant to the topics of interest. Many valuable specialized databases appear to be underutilized (Hightower, 1998). This conclusion has been borne out by recent transaction log analyses done at UIUC.

Libraries are making concerted efforts to help point users to the best resources, but the most common practices are becoming less effective and more difficult to maintain as the diversity and sophistication of available resources continues to increase rapidly. Most libraries now provide on their website at least some organization of available resources. The current UIUC Gateway provides 4 views of campus-wide online journal article index resources: listed alphabetically by Title; grouped by vendor; categorized under 8 subject headings (see **Figure 1**); and segregated according to whether they include links to article full-text. Many subject specialist librarians also generate selective lists of resources particularly useful to their area of specialization. Users also can get help from finding aids available in print and/or over the Web and from online tutorials provided by some of units. The end result however is that users generally must pick from static lists, relying only on resource titles (augmented in some cases with 1 or 2 sentence short descriptions) to make their selection of which resource to search.

The UIUC Smart Database Selector Project

In 1998 preliminary research was done examining alternative methods of helping users better identify the information resources most likely to assist them in their search and reviewing work being done elsewhere in this area. With funding from the UIUC Research Board and the UIUC Library's Research and Publication Committee, a project was undertaken in 1999 to create a prototype of a Web-based expert system tool that

could assist users in identifying the most appropriate online or CD-ROM database to meet a particular information need. The project objectives:

1. to identify and better understand computer-assisted techniques for guiding users to the databases most suitable for their particular information needs;
2. to pioneer innovative methods to improve user services given the proliferation of electronic databases and digital information within individual institutions; and,
3. to improve the ease of access to information by taking advantage of the Web and related protocols and technologies.

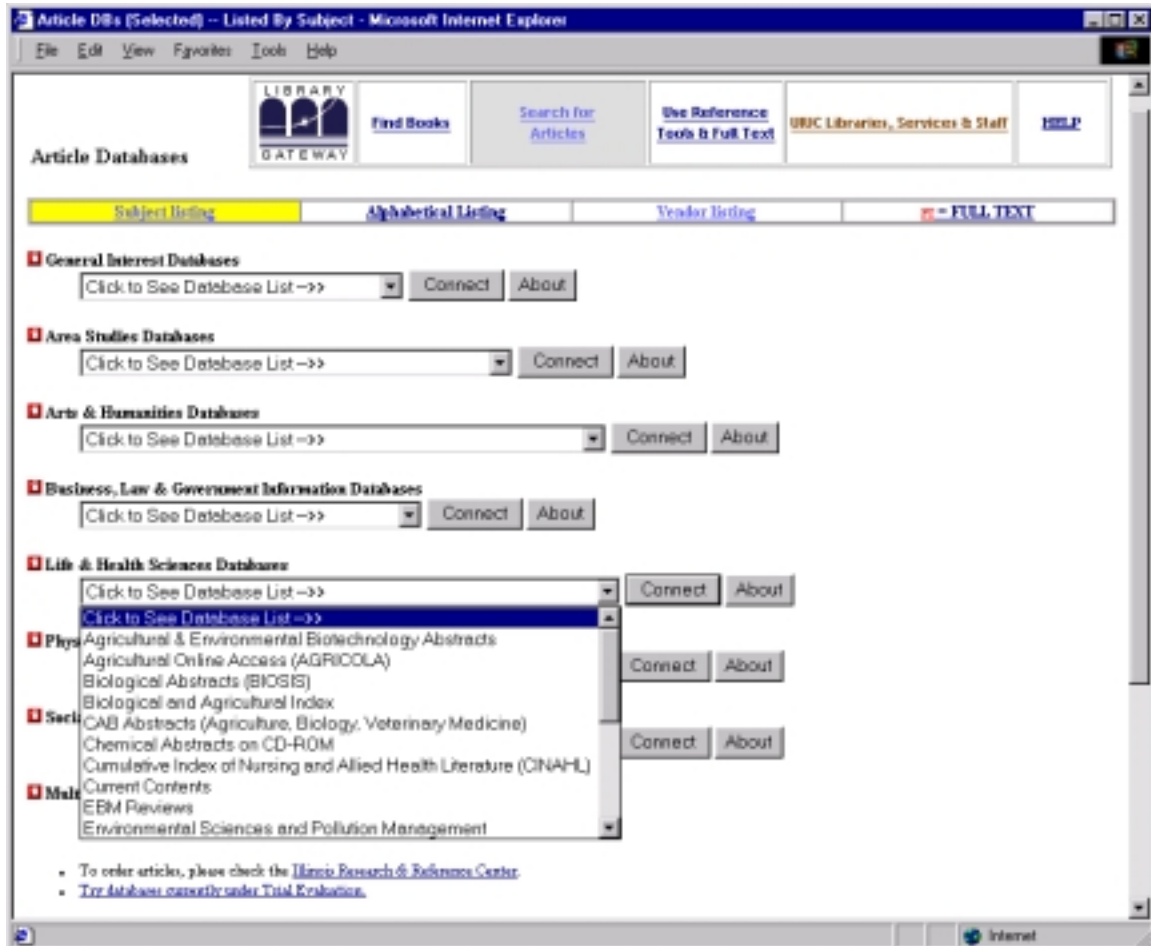


Figure 1 – Article Index Database Pick List Organized by Subject

A prototype system designed to assist users in selecting from among the article databases, directories, statistical and government document electronic resources presently available at UIUC has been created and is being reviewed by members of the Library faculty. Limited end-user testing and focus group evaluations also are in progress.

Using this prototype tool, users are able to select resources by any of 3 methods: 1) searching for relevant resources using free-text keywords and phrases; 2) browsing

available databases by subject categories; or 3) choosing desired database by characteristics only (e.g., material types indexed, chronological coverage, etc.). **Figure 2** shows query entry screen for this tool. Several filters can be used to narrow keyword and browse selections.

The screenshot shows the 'Smart Database Selector' web interface. At the top, it says 'Use one of these forms to suggest or help select the database(s) most suited to your information need.' There are three main sections:

- Select by Keyword(s)** (Type in keyword(s) from your topic; also you may specify other criteria that apply): Includes a 'Keyword(s):' text box, radio buttons for 'and' and 'or', and 'Select' and 'Reset' buttons. Below are 'Other Criteria (optional)' with dropdowns for 'Choose a Material Type', 'Choose Content', 'Choose a Database Type', 'Choose a Broad Subject Area', 'Choose database level', and 'Choose Languages'.
- Browse List of Database Topics** (You may limit your browse by specifying additional criteria that apply): Includes a 'Click a letter to browse' field with a list of letters (A-Z), a 'Reset' button, and the same 'Other Criteria' dropdowns as above.
- Select by Material Types and/or Content** (Choose content and/or material type; then specify other criteria): Includes the same 'Other Criteria' dropdowns, a 'Type in a time period' field with 'After' and 'Before' sub-fields, and 'Select' and 'Reset' buttons.

At the bottom, there is a link 'About the project'.

Figure 2 – User Interface of Expert System Database Selection Tool

The approach we've taken varies from earlier and ongoing library projects elsewhere – e.g., EasyNet (Hu, 1987, 1988; McCarthy, 1986; O'Leary, 1988; van Brakel, 1988), DialIndex / DialogWeb, UCSD's "Database Advisor" (Hightower 1998) – in that it does not involve simultaneous, real-time searching of multiple resources, which enables inclusion of a diverse collection of Online and stand-alone electronic databases, such as CD-ROM databases and license restricted access databases. It also differs from generic Web search engines – where the emphasis lately has been on query formulation, interface design, and results display (Feldman 1998) – in that more extensive filtering is possible and automatic indexing of resources is augmented with manually assigned characteristics.

The approach we've taken is to create a selection database containing rich and extensive characterizations of the available resources. User queries and

characterization filters are then applied against that selection database to come up with a ranked list of resources of possible interest. As described in more detail below, the process is intended to mimic the process a librarian uses when selecting a database to search for a particular information need.

Resource characterizations were developed using database documentation, standard reference sources, information supplied by librarians most knowledgeable about the subject area of each resource, and whenever available, the controlled vocabularies of the resources themselves. A standard form was created and terminology defined to assist in the gathering and assigning of resource characteristics. Controlled vocabulary information was obtained using 2 techniques: 1) direct processing of the complete controlled vocabulary as supplied by the database producer; or 2) by large-scale random sampling of recent database records. The latter approach had the advantages of being more highly automated and of returning a selection of controlled vocabulary terms biased towards the most heavily used terms.

At this time, one advantage of the approach taken is that all the data used to assist in resource selection resides on local systems. The process is not dependent on primary data sources being always available and therefore tends to scale better in the current environment (the prototype selection database contains information for about 150 resources with controlled vocabulary data for more than half that number). The biggest drawback of this approach is that the thorough characterization of the resources takes considerable manual labor, both initially when a resource is first added to the selection database and then (to a lesser extent) on an ongoing basis to maintain resource entries.

Resource Selection Algorithm

The process of actually identifying relevant resources based on the characterization information stored in the selection database was designed to mimic the way sophisticated end-users and librarians accomplish this task. The search patterns of users seeking information were analyzed and the strategies which reference librarians use to respond to users' information requests were studied. A database selection algorithm, which attempts to replicate the professional librarians' database selection strategy, was developed in the summer of 1998 from the above studies, and from the discussions and inputs of several UIUC librarians consulted.

Resource Selection Algorithm:

Topic/subject category term +/- broad subject +/- format +/- content +/- level +/- database type => database

This formula states: Topic keywords or specific subject terms combined with broad subject area, format of information sought, type of content sought, level of information need, and/or database type will determine the databases relevant to the topic.

The following process demonstrates how a librarian's decision-making model is translated to an algorithm for database selection:

1. **Topic keywords:** A user who needs to search electronic databases should have a topic or specific subject to research. He/she can often express that topic

in appropriate natural language. For example, a user would express his or her topic: “The position of women and the attitudes towards women in Hispanic American community.” Nouns and noun phrases (e.g., “women” and “Hispanic American community”) can then be extracted and used to help identify appropriate information resources that should be consulted.

- 2. Subject category terms:** If the user has difficulty expressing a topic in natural language terms, a browse list of detailed subject terms assigned to resources by knowledgeable librarians is an appropriate alternative.
- 3. Broad subject:** When a librarian is approached for assistance, the librarian usually first tries to determine the broad subject area of interest to the user. For example, the topic “cloning” can have many aspects –technical/medical (how to clone), social effects of cloning, legal aspects of cloning, etc.
- 4. Format:** The librarian then might determine what format(s) the patron is interested in – e.g., newspaper articles, magazine articles, journal articles, conference proceedings, government documents, patents, dissertations, trade publication, etc.
- 5. Contents:** The librarian might also try to find out what type of materials are of most interest to the patron – e.g., biographical information, current events, historical information, statistical information, etc.
- 6. Level:** Often the librarian also will try to characterize the research level of the patron’s interest. Is the patron wanting to do a brief paper for Freshman rhetoric, a senior class design project, initial research for a thesis, research for a possible article publication or conference presentation, etc.
- 7. Database type:** Based on the information obtained from the reference interview and knowledge of what is available in the library, the librarian may also make an assessment as to which type of databases is most suitable, e.g., bibliographic, full-text, directory, or statistical.
- 8. Database:** Combining all or some subset of the filters and search criteria described in the proceeding steps, the librarian then will be able to point the patron to the resources most likely to help meet the information need.

Building the Prototype System

Having defined the approach to be used, gathered the characterization data required, and developed a model of the selection decision process, what remained was to construct a prototype version of the system. An obvious requirement was that the system needed to be Web-based. A simple, 3-tier Web application architecture was defined and is depicted in **Figure 3**. Hardware used was a dual-Pentium II server. Microsoft Internet Information Server (ver. 4) and SQL Server (ver. 7) were selected for the application backend.

Where feasible complete controlled vocabularies were obtained in ASCII format from the database producers. Any pre-processing of these files (e.g., to eliminate unnecessary formatting and headings) was done using customized VBScript modules. The files were

then uploaded into SQL tables. In some cases where digital versions of controlled vocabularies were not readily available, we were able to capture controlled vocabulary terms directly from database records. Typically 10 to 20 % of the records for the most recent complete calendar year available were sampled, and controlled vocabulary terms were harvested. This process was accomplished using a VBScript module that generated the appropriate Z39.50 protocol queries, processed the results returned, and then loaded the terms captured into SQL tables. For a few of the resources that lack formal controlled vocabularies (e.g., Current Contents) a similar process was used to sample other fields (e.g., author-assigned keywords).

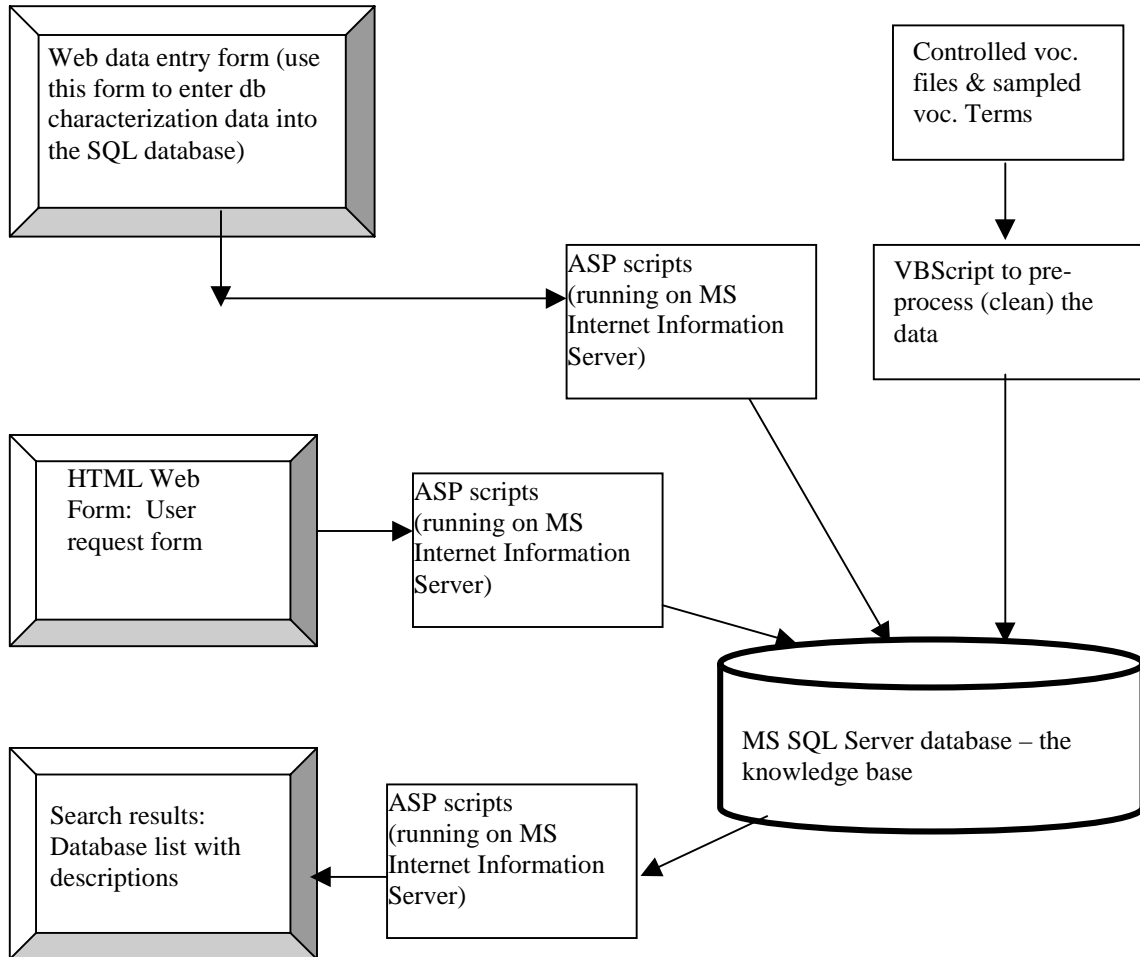


Figure 3. The Architecture (components) of the prototype system

The selector database was designed as a multi-table, relational database. A master record was generated for each resource and was stored in the primary table. Characterization data used for basic filtering (e.g., Broad Subject, Format, Database Contents, Level, and Database Type) were stored in discrete tables with linking or pivot tables used to link such characteristics to each master record. Assigned subject category terms and controlled vocabulary terms were stored in separate tables, and Microsoft SQL 7 full-text catalogs of the fields in these tables were made. These full-text catalogs are essentially indexes automatically created by the database application by

analyzing the terms in the tables. Noun phrases are extracted from the terms and indexed. Stopwords are discarded. The indexes are optimized for efficient keyword and phrase searching. This approach obviated the need to manually build our own special keyword indexes or rotated term indexes from the controlled vocabulary and subject category data. The current controlled vocabulary full-text catalog was built after analysis of 2.4 million different terms consisting of 690,000 unique words. Search performance using this built-in feature of Microsoft SQL 7 in this way has been quite good.

The SQL query structure used is moderately complex. Right-hand truncation is used in conjunction with the built-in full-text searching features alluded to above. A crude ranking of databases satisfying the query criteria is done using the frequency of term matches as the figure of merit. This tends to under rank databases for which there are no controlled vocabulary entries. Further investigation of how to minimize any bias against databases for which controlled vocabulary can't be obtained remains to be done.

Remaining Issues

The work done in designing and creating this prototype tool indicates that the approach has promise. There remains a great deal of evaluation work to be done to better quantify that promise, and there remain a number of issues to resolve before the prototype could be upgraded to a more production-ready utility.

In terms of evaluation, it's still not clear how useful librarian-assigned terms and classification schemes, even in combination with database controlled vocabulary, really are in resolving end-user queries. Whether this approach might offer better or comparable recall and precision to schemes that simultaneously search multiple databases remains to be seen. By soliciting user feedback and monitoring how users make use of the prototype tool, we hope to shed further light on this issue.

There also remain major cost and production implementation issues. Based on our experience the initial investment to properly characterize resources is high. However, this task could be eased and better distributed by having the library's subject specialists more involved in the database characterization process. It's not yet clear whether the cost to maintain this information is also high. That will depend on how rapidly major and fundamental changes are made to individual databases. Re-sampling of database controlled vocabulary could be largely automated to keep up with changes in controlled vocabulary and the use of that controlled vocabulary. Frequent changes in database scope, format, etc., would be more costly to keep up with, but currently such changes are less common. Finally, there would need to be some measure of cooperation on the part of database producers. Producers have legitimate intellectual property claims with regard to controlled vocabulary and also may have concerns about how their databases are otherwise characterized in systems such as the one investigated. Reluctance to provide information about their databases, or severe constraints on how such information may be used would inhibit development of this kind of system. Better consistency in the information provided by the various database producers about the resources they make available would facilitate development of this kind of system.

References

Feldman, S. (1998), "Web search services in 1998: Trends and challenges", *Searcher*, Vol. 6 No. 6 pp. 29-39.

Hightower, C, Reiswig, J and Berteaux, S. S. (1998), "Introducing Database Advisor: A new service that will make your research easier", *College and Research Libraries News*, Vol. 59, No. 6, pp. 409-12.

Hu, C. (1988), "An evaluation of a gateway system for automated online database selection", In: *Proceedings of the Ninth National Online Meeting*. Medford, NJ, Learned Information, pp. 107-114.

Hu, C. (1987), "An Evaluation of an Online Database Selection by a Gateway System with Artificial Intelligence Techniques", *Doctoral dissertation*. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science.

McCarthy, M. V. (1986), "InfoMaster: a powerful information retrieval service for business", *Online*, Vol,10, No. 6, pp. 53-58.

O'Leary, M. (1988), "Easynet revisited: pushing the online frontier", *Online*, Vol. 12, No. 5, pp. 22-30.

Thornburg, G. E. (1987), *LOOK: Implementation of an Expert System in Information Retrieval for Database Selection. Doctoral dissertation*. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science.

van Brakel, P. A. (1988), "EasyNet: intelligent gateway to online searching", *South African Journal of Library and Information Science*, Vol. 56, pp. 191-197.

Zahir, S. and Chang, C. L. (1992), "Online-Expert: an expert system for online database selection", *Journal of the American Society for Information Science*, Vol. 43, pp. 230-357.