Interface Design Issues in Providing Access to Databases Using Semantic Classification Techniques

Philip J. Smith
Bekka Denning
Cognitive Systems Engineering Laboratory
Department of Industrial and Systems Engineering
The Ohio State University
Columbus OH 43210

INTRODUCTION

A primary goal of classification is to assist information seekers in finding materials of use to them. These materials may be books, they may be songs, they may be examples of architectural designs.

Given this goal, it is not sufficient to ask:

> How can we classify a set of materials so as to discriminate among the members of this set?

Instead, we must ask the broader question:

> What classification will be most effective in helping information seekers to achieve their goals?

To answer this question, classification research becomes closely intertwined with questions about the needs of prospective information seekers, about the ways in which they can access and use the results of the classification process, and about the cognitive processes of the information seeker that arise during the search process. In this paper, we explore some of the issues raised by this broader systems perspective.

Information Seeking as a Problem of Learning about a Topic

In an empirical study of the interactions between information seekers and expert intermediaries while searching a bibliographic database, we found that much of the activity focused on exploring a topic area in order to define the specific topic of interest (Smith, Shute, Galdes and Chignell, 1989). Very few information seekers came to the search with fixed, well- defined topics. Rather, they came with some broader goal, such as:

> I need to write a term paper for a course on environmental chemistry. I'd like to focus on acid rain.

Thus, one goal that classification efforts need to support is that of topic definition.

Semantic Classification

If there is a good match between a semantic classification scheme and the needs of the information seeker, this approach offers a powerful tool to support this learning and exploration process. Using a frame-based representation to classify document contents, for instance, it is possible to explicitly support exploration in an abstraction space. Instead of exploring by browsing retrievals on individual documents, the user can browse through the set of available slots (drug, use for drug, source of drug) for categories of interest, and can browse through the semantic primitives or slot-fillers associated with each slot. Furthermore, because the computer understands the relationships among these slot- fillers, and understands the contents of individual documents at a semantic level, the computer can provide tools to assist the user when browsing in this abstraction space. The computer can, for instance, help the user focus attention by listing for her those side-effects known to result from the use of a particular drug such as protriptxline. (Additional details on such aids are given later in the paper.)

Consider, for example, a student writing a paper for a course on pharmacology. Using a typical character-string search system, she would have to generate queries such as

PHARMACOL? OR DRUG#

and look at the resultant retrievals in order to generate a more specific topic. Since such retrievals are not organized in any fashion, this is a very time-consuming, hit-or-miss approach to exploring the subject area. If relationships among the index terms have not been specified, the same problems arise even if a controlled vocabulary has been used to index documents.

Using a frame-based search system (Humphrey, 1989; Monarch and Carbonell, 1987; Smith and Chignell, 1984; Vickery and Brooks, 1987), however, it is possible to develop an interface that presents the user with an abstract description of the subject area. An example of this is provided by ELSA (Electronic Library Search Assistant), a prototype system that we have developed.

An Interface to a Frame-Based Classification System.

ELSA is designed to help in searches concerned with pharmacology. The search engine for this system assumes documents have been indexed as frame instantiations.

To assist information seekers in defining and exploring their topics, ELSA explicitly displays the concept categories or slots used to index documents (see the window on the left in Figure 1). Thus, ELSA indicates that the documents in its database have been indexed in terms of the drug discussed, its uses, its side-effects, etc.

ELSA also provides the information seeker with a powerful tool for exploring specific subtopics. In trying to define her topic our pharmacology student might, for instance, look at the screen displayed in Figure 1 and say:

That's interesting. Maybe I should focus on drugs used for a particular purpose.

She might then ask:

> I wonder what different uses of drugs I could consider focusing on in my paper?

The window on the right in Figure 2 answers this question. Because the cursor is in the "Use for Drug" box in the left window, the right window lists all of the uses of drugs included in the indexing of the available documents. Figure 2 shows the student initially selecting narcotics for exploration

This student might then ask:

> Maybe I should focus instead on the use of a particular drug as an antidepressant. I wonder what drugs are possible?

Because of its frame-based indexing, ELSA can also assist in answering this question. If the student begins her search by entering antidepressant as the use of interest by either typing it in or by clicking on that item in the list of "Available Uses for Drugs", she can then view the list of relevant drugs. This is accomplished by next moving the cursor to the "Drug:" box (indicating that she wants to see the list of "Available Drugs," and then clicking on the "Show Pruned List" button. This results in a display of only those drugs discussed in the database as having usefulness as antidepressants (see Figure 3).

Thus, users of ELSA can view lists of the available slots (drug, use for drug, etc.) and slot-fillers (semantic primitives such as trazodone as a drug, or antidepressant as a use for drugs). This explicit presentation helps the user to understand the organization of ELSA's knowledge-base, and to search for concepts of interest by browsing through the lists of slots and slot-fillers. While this explicit display of the semantic classification scheme solves some traditional problems in searching bibliographic databases, it potentially creates new ones. Suppose, for instance, that the information seeker entered antidepressant in the "Drug" box instead of the "Use for Drug" box. ELSA would display the list of "Available Drugs" and indicate that there was no match for antidepressant. Preliminary tests of ELSA have indicated that, without assistance, this is a common error that is difficult to recover from.

This suggests a potential problem with classification schemes. The imposed organization may help the user to learn about a topic area by browsing the displayed lists of slots and slot-filler but, if she already knows what she is looking for, she may look in the wrong place and consequently conclude it is unavailable.

Based on observation of such problems, ELSA has been enhanced so that, if no match is found for antidepressant entered as a drug, ELSA searches the other slots for a match. In this case, ELSA indicates to the user that information is available on the use of certain drugs as antidepressants, and ask if this is what they really wanted to indicate when they typed in antidepressant as a drug of interest

Figure 1. Initial Dispaly for ELSA.

Figure 2. Display of the Different Uses of Drugs Discussed in the Database (Their Use as Narcotics has been selected.)

Figure 3. A Display Answering the Question: What Drugs are Used as Antidepressants?

Information Seeking as a Problem of Labelling

Whether the user comes in with a well-defined topic or develops one following some exploration, another classic problem that classification research deals with is labeling. Because of its flexibility, natural language offers many ways to express the same concept.

The use of controlled vocabularies for indexing creates equivalence classes such that, if the user finds the appropriate controlled vocabulary term, she is assured that she will retrieve all of the relevant documents regardless of how the original author expressed herself (not counting indexing errors, of course). This solution raises a new problem, though:

> How can we ensure that the user will find the right controlled vocabulary term?

This same problem arises when documents are indexed as frames. In this type of classification scheme, the semantic primitives used as slot-fillers (e.g., desipramine as a drug or antihistamine as a use for certain drugs) are in essence controlled vocabulary terms (Smith, et al., 1989). The solution taken to solve this problem by ELSA and its predecessor EP-X is fairly standard. Each semantic primitive has an associated set of triggers, including cognates and synonyms.

Frame based classification schemes do raise one unique issue, however, regarding the labeling problem. It is possible that the same semantic concept can be a slot-filler in two different slots, playing a much different role in each one. For example, in EP-X, a user expressing an interest in the chemical mercury is asked whether she is interested in mercury as a pollutant or as a component in a mercury vapor lamp (equipment which is used to treat biological pollutants).

Information Seeking as a Problem of Generating Class Members

The use of triggers to deal with cognates, synonyms and role ambiguities resolves the problem of labeling by creating appropriate equivalent classes. There is, however, another type of equivalence class that semantically-based systems must deal with, resulting from IS-A-KIND-OF and IS-A-PART-OF relationships.

When a user expresses interest in a concept like Ohio or natural waters, she generally wants to include specific cases of this concept. This raises a challenging question:

> How do we help the user to explore the contents of hierarchies and to easily indicate her interests?

ELSA does not yet deal with this issue. In its predecessor EP-X, though, a number of functions were provided to support exploration of hierarchies. As Figure 4 illustrates, users could:

1. Display all of the children of a concept;
2. Display only those children relevant to the current topic;
3. Delete or undelete individual children in the hierarchy. (When a child is deleted, an EXCEPT operator is used instead of NOT.)
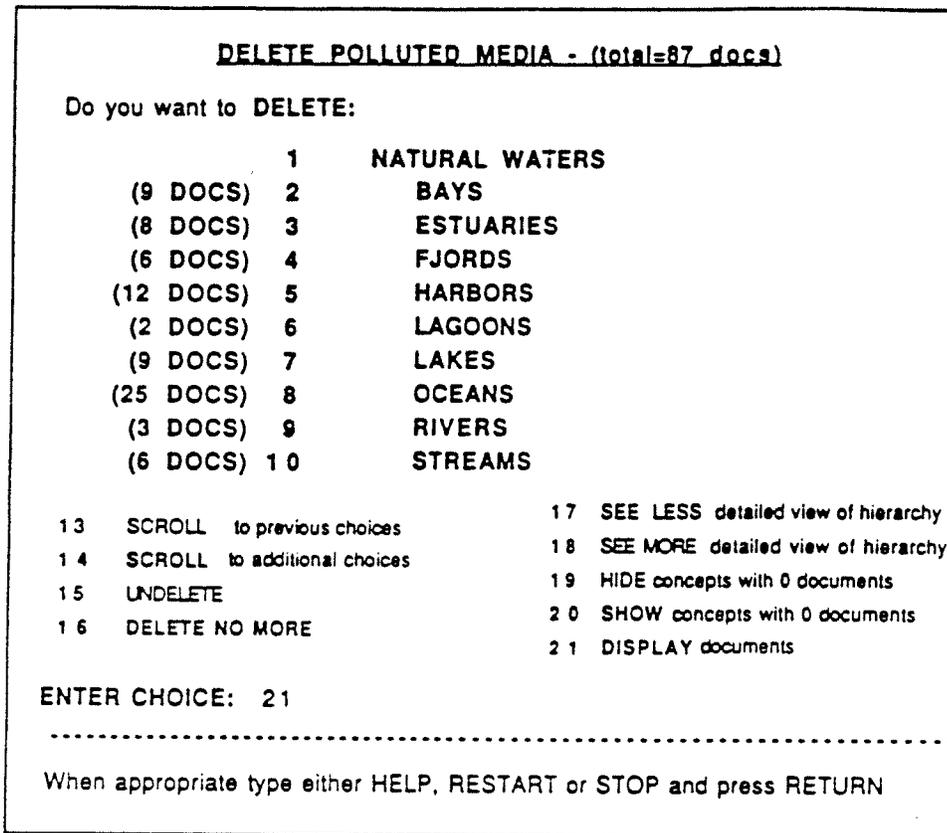4. View the hierarchy at different levels of detail.

145

```
          DELETE POLLUTED MEDIA - (total=87 docs)

Do you want to DELETE:

                        1      NATURAL WATERS
            (9 DOCS)    2         BAYS
            (8 DOCS)    3         ESTUARIES
            (6 DOCS)    4         FJORDS
           (12 DOCS)    5         HARBORS
            (2 DOCS)    6         LAGOONS
            (9 DOCS)    7         LAKES
           (25 DOCS)    8         OCEANS
            (3 DOCS)    9         RIVERS
            (6 DOCS)   1 0        STREAMS


   1 3   SCROLL   to previous choices      1 7   SEE LESS detailed view of hierarchy

   1 4   SCROLL   to additional choices    1 8   SEE MORE detailed view of hierarchy

   1 5   UNDELETE                          1 9   HIDE concepts with 0 documents

   1 6   DELETE NO MORE                    2 0   SHOW concepts with 0 documents

                                           2 1   DISPLAY documents


   ENTER CHOICE:   21

   ........................................................................

   When appropriate type either HELP, RESTART or STOP and press RETURN
```

Figure 4. Display of an Available Hierarchy in EP-X (From Smith, Shute, Chignell and Krawczak, 1989).

This solution provides the user with a powerful tool for exploring a hierarchy. It leaves unresolved, however, a very difficult question:

How should tangled hierarchies be displayed and navigated?

Our current feeling is that this issue is best dealt with by creating hierarchy clusters that are small enough to present in their entireties, with tanglings eliminated by representing alternative organizing perspectives as separate clusters. When the searcher then displays the hierarchy information for a concept such as lead-210, its memberships in clusters organizing radioactive substances and organizing heavy metals can both be displayed.

Information Seeking as a Problem of Reasoning about Temporal Sequences

Synonyms and cognates as triggers, semantic primitives as controlled vocabulary terms, hierarchical clusters as representations of IS-A-KIND of relationships - these are all well within the bounds of traditional classification research. As discussed above, there are some interesting issues regarding how to provide access to such knowledge in support of document retrieval, especially if we go a step beyond thesauri and begin using frames as the underlying knowledge structure.

Our research on the cognitive processes of expert intermediaries indicates, though, that there is another class of knowledge that can be used to form equivalence classes. This knowledge is concerned with temporal (or spatial) sequences.

Consider the following example. An information seeker expresses an interest in documents on the control of acid rain. The intermediary (who is knowledgeable about the literature on environmental chemistry) responds that they should go beyond a query such as:

CONTROL?AND ACID?(W)RAIN

and focus on documents dealing with the prevention of air pollution by sulfur and nitrogen oxides. Her reasoning is that:

1. The information seeker is looking for ways to control acid rain;
2. Pollutants pass through a number of stages over their life histories;
3. It is easier to control different pollutants at different points in their life history. In particular, pollutants can be controlled before they are formed (by changing the originating process), before they are released into the environment, and at various points in their movement through the environment;
4. .In the case of acid rain, it is easier to control it by controlling the formation and emission of its precursors (sulfur and nitrogen oxides in gaseous form). Therefore, documents dealing with this part of the life history need to be retrieved.

This reasoning *process* is thus the basis for classifying two topics as related. Similar processes are easily identified in other fields as well, such as medicine. Thus, to meet the needs of the users more fully, we need an interface that provides access to such process knowledge, and we need to develop classification methods to represent it. This requires classification methods that capture relationships among topics rather than simply relationships among individual concepts. It further requires an interface that supports exploration of the relationships among topics.

CONCLUSION

The general theme of this paper is that, to develop useful classification schemes, we need to know how and why these classifications will be used. This question becomes particularly intriguing when we start to introduce |intelligent| functions into computer systems that make use of semantic classification schemes. Such new tools may solve certain problems. They may, however, create new ones if the classifications and supporting search functions are not carefully designed.

147

ACKNOWLEDGEMENTS

REFERENCES

1. Humphrey, S. M. (1989) MedIndEx System: Medical indexing expert system. *Information Processing and Management*, 25, 73-88.

2. Monarch, L., and Carbonell, J. CoalSORT: A knowledge-based interface. *IEEE Expert*, (1987) 39-53.

3. Smith, P. J., Shute, S. J., Chignell, M. H., and Krawczak, D. A. "Bibliographic Information Retrieval: Developing Semantically- Based Search Systems". *Advances in Man-Machine Systems Research,* Vol. 5, pages 93-152.

4. Smith, P. J., Shute, S. J., Galdes, D., and Chignell, M. H. "Knowledge- based search tactics for an intelligent intermediary system". *ACM Transactions on Information Systems,* Vol. 7, No. 3, July 1989, pages 246-270.

5. Smith, P. J., and Chignell, M. Development of an expert system to aid in searches of the *Chemical Abstracts*. In *Proceedings* of the 47th ASIS Annual Meeting (Philadelphia, Pa., Oct. 21-25). Knowledge Industry Publications, White Plains, 1984, pp. 99-102.

6. Smith, P. J., Shute, S. J., Chignell, M. H., and Krawczak, D. Bibliographic information retrieval: Developing semantically-based search systems. In *Advances in Man-Machine Systems Research*, Vol. 5, W. Rouse, Ed. JAI Press, Greenwich, Conn., 1989, pp. 93-117.

7. Vickery, A., and Brooks, H. M. PLEXUS--the expert system for referral. *Information Processing Manage.* 23, 2(1987), 99-117.