Smith, P. J., Denning, R., Shute, S. J., & Normore, L. F. (1990). Toward the development of semantically-based search systems. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 155-160.
doi: 10.7152/acro.v1i1.12475

Toward the Development of Semantically-Based Search Systems

Philip J. Smith and Rebecca Denning

Cognitive Systems Engineering Laboratory, The Ohio State University
Columbus, OH 43210, USA

Steven J. Shute

AT&T Bell Laboratories, Holmdel, NJ 07733

Lorraine F. Normore

Chemical Abstracts Service, Columbus, OH 43210

Over the past several years, we have conducted a number of empirical studies focusing on the performance of human search intermediaries and indexers. Based on insights from these studies, we have developed a computerized intermediary system (EP-X) that represents document contents as frames, and that uses knowledge-based search tactics to assist information seekers in exploring the contents of such frame-based document databases. Below, we present for discussion several propositions based on our experiences in studying human experts and in building computerized intermediaries.

Proposition 1. One of the goals of indexing is to create an abstract representation of the contents of documents. The purpose of this abstraction is three-fold:

A. To allow an information seeker to judge the relevance of an individual document based on its document surrogate (containing the author, title, abstact, and indexing entries);

B. To provide an efficient, effective way of accessing a cluster of documents relevant to some topic;

C. To support exploration of this abstraction space, so that the information seeker can determine which clusters of documents she wishes to access.

As an example, consider the abstract representation of one document in the *Chemical Abstracts* (from Smith, Shute, Chignell and Krawczak, 1989):

IT Environment Pollution
(By mercury, in Japan, Minimata disease of humans in relation to)

IT Mental Disorder
(Retardation, of children, of Japan, methylmercury in relation to)

IT 7439-97-6, Biological Studies
(Environmental pollution by, in Japan, Minimata disease in humans in relation to)

IT 7439-97-6W, Me Compds
(Of children, of Japan, environmental pollution in relation to)

Based on such indexing, an information seeker can easily determine that this document discusses mercury as a pollutant in Japan and its relationship to Minimata disease. For an information seeker interested in "the effects of exposure to mercury as a pollutant," this

ISSN: 2324-9773

Smith, P. J., Denning, R., Shute, S. J., & Normore, L. F. (1990). Toward the development of semantically-based search systems. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 155-160. doi: 10.7152/acro.v1i1.12475

abstraction serves well the purpose of allowing her to make an accurate judgment of relevance. For a person interested in "the mechanisms underlying mercury's neurotoxic effects," however, this abstraction is less adequate. If does not provide sufficient detail to know whether the document discusses the mechanisms of neurotoxicity. Thus, one critical problem in designing an indexing scheme is to identify the appropriate levels of detail to meet the needs of different information seekers.

Current indexing practices for the *Chemical Abstracts*, like many other bibliographic databases, run into greater difficulties, however, in meeting the second goal of this abstraction space, to provide efficient, effective access to clusters of related documents. These problems arise because access is achieved via queries consisting of character strings connected by logical operators. Such problems, well recognized in the literature on information retrieval (Salton and McGill, 1983), include the following:

A. Identifying appropriate synonyms and related terms (e.g., pollutants, wastes, effluents) to retrieve documents relevant to some semantic concept;

B. Specifying terms to retrieve documents on all the specific cases of some general concept of interest (e.g., including the names of specific countries, cities, lakes, etc. to retrieve documents on the effects of acid rain in Western Europe);

C. Dealing with the ambiguities of lexical terms (e.g., that the query POLLUTION AND PRECIPITATION will retrieve documents on acid rain, but also documents on the use of precipitation as a method to remove pollutants from wastewater);

D. Determining the appropriate set of logical operators to use.

Perhaps the greatest difficulties arise, however, in terms of the third purpose in creating such an abstraction, that of supporting interactive exploration to find relevant clusters of documents. In most databases, indexing is done at a very specific level. If there is no good conceptual map to relate these specific entries to each other or to broader concepts, it is a very challenging task to explore this abstract space in search of relevant documents. As an example, consider a person interestd in documents on processes used to prevent pollution from cadmium. To conduct a search that has high recall and good precision, this person would have to know the names of all of the specific removal processes and enter them in her query. Similarly, a person interested in pollution from mercury would have to include all relevant mercury compounds (such as methylmercury) in order to assure high recall.

Proposition 2. The explicit representation of document contents as frames, with hierarchically-defined semantic primitives as the slot-fillers in these frames, offers an alternative abstraction that may overcome many of these difficulties (Smith, Shute, Galdes, and Chignell, 1989). By representing documents in terms of their meanings, and by building an explicit conceptual map representing the semantic relationships among documents, many of the shortcomings of character-string searches may be avoided.

As an illustration, consider the semantic primitive *Removal Processes*. In our frame system, this primitive is hierarchically defined in the sense that the computer knows that there exist two classes of removal processes (biological and chemical/physical), and that each of these classes

ISSN: 2324-9773

Smith, P. J., Denning, R., Shute, S. J., & Normore, L. F. (1990). Toward the development of semantically-based search systems. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 155-160. doi: 10.7152/acro.v1i1.12475

includes a variety of more specific processes (such as adsorption, chelation and ion exchange). Such knowledge (which must be provided to the system by human experts) can be used for several purposes. Consider again the person who is interested in processes to prevent pollution from cadmium. She could enter a list of keyword phrases, such as

CADMIUM
REMOVAL PROCESSES

(or some other triggering phrases such as *CD* or *PREVENTION*). The system automatically retrieves documents discussing any of the processes used to remove cadmium from wastes. The list of relevant processes can then be displayed and associated document information viewed.

Hierarchical semantic primitives are slot-fillers in specific frames. One such frame, labeled *Removal of Pollutants* contains several slots such as *Pollutant*, *Polluted Medium* and *Removal Processes*. When an information seeker expresses an interest in a topic such as "the removal of cadmium from wastewater," the computer can use the structure of this frame to suggest ways to refine the topic if desired. In particular, accessing the *Removal Process* slot actually activates another frame labeled *Removal Processes*, which has three slots of its own (*Specific Removal Processes*, *Equipment used in Removal*, and *Chemicals used in Removal*). Using knowledge contained in instantiations of this frame, the computer can generate suggestions such as:

To narrow your topic, you could limit your search to specific removal processes (such as chelation or ion exchange).

Furthermore, the computer can list the set of removal processes relevant to the original topic (the removal of cadmium from wastewater) if the information seeker wants to explore this direction for narrowing her topic.

Proposition 3. Given the existence of such a frame-based database, powerful knowledge-based search tactics can be applied by a computerized intermediary system to assist information seekers in defining and refining their topics of interest. These tactics mimic the types of assistance provided by human intermediaries (Shute, 1989).

The types of knowledge illustrated in Proposition 2 can be further utilized by the system to generate suggestions for related topics. As a second example, consider a person asking for documents on "the removal of cadmium by chelation." This request retrieves very few documents. The system can be helpful, however, by suggesting alternative removal processes (such as ion exchange) that are discussed in the available documents as being used to remove cadmium from wastes.

Proposition 4. The contents of document surrogates in the *Chemical Abstracts* are well described in terms of frames. The current indexing of many documents is highly stereotypic. In developing our frame system for the field of environmental chemistry, for instance, we have found that a total of 47 slots is sufficient to capture the contents of indexing at Chemical Abstracts Service (CAS) on this topic. Less rigorous analyses for the fields of pharmacognosy, pharmacology, and biotechnology indicate that comparable totals exist for other topic areas.

Proposition 5. The intellectual process of indexing at CAS is well modeled in terms of the activation and use of frames or schemata in the minds of indexers (Normore and Smith, 1989;

ISSN: 2324-9773

Smith, P. J., Denning, R., Shute, S. J., & Normore, L. F. (1990). Toward the development of semantically-based search systems. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 155-160.
doi: 10.7152/acro.v1i1.12475

Smith, Chignell, and Krawczak, 1984). Thus, the intellectual work required to produce a frame-based database is already part of the intellectual process. The results of this intellectual process, however, are currently represented in the computer in a form such that much of the indexer's knowledge about a document's meaning is unavailable to the computer. As a result, the computer cannot use this knowledge to assist information seekers in finding relevant documents.

Proposition 6. Given the development and use of a suitable indexing assistant (Humphrey, 1989), the results of indexers' intellectual efforts could be represented as frames, thus avoiding the current loss of knowledge resulting from existing indexing procedures. Furthermore, such an indexing assistant might significantly reduce training and production costs for creating a database such as the *Chemical Abstracts*, and might also improve consistency in the database (Normore and Smith, 1989).

CONCLUSION

These propositions lead toward the conclusion that, for databases like the *Chemical Abstracts*, where substantial indexing by human experts is already the practice, there are substantial arguments for developing frame-based indexing and search assistants. Such computer aids offer the potential to reduce costs in the indexing process and to greatly improve performance by information seekers.

Two cautions are appropriate, however. First, frame-based search intermediary and indexing assistants have not yet been subject to extensive effectiveness or feasibility testing. Current arguments for such systems are based more on face-validity than on performance data. Second, there are many interface design issues that arise, and that could potentially confound efforts to evaluate them. Thus, there remains considerable research to be done before we can adequately judge the validity of the conclusion toward which these Propositions lead us.

REFERENCES

Humphrey, S. (1989) A knowledge-based expert system for computer-assisted indexing. *IEEE Expert*, 25-40.

Normore, L. and Smith, P. J. (1989) *An Analysis of Quality Control Reviews*, CAS Technical Report, Chemical Abstracts Service, Columbus, OH.

Salton, G. and McGill, M. (1983) *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Shute, S. (1989) *An Empirical Investigation of Knowledge-Based Search Tactics in the Topic Refinement Behavior of Online Bibliographic Searchers*. Ph. D. Dissertation, The Ohio State University, Columbus, OH.

Smith, P. J., Chignell, M., and Krawczak, D. A. (1984) Development of a knowledge-based bibliographic information retrieval system. *Proceedings of the 1984 IEEE Conference on Systems, Man and Cybernetics*, 222-225.

ISSN: 2324-9773

Smith, P. J., Denning, R., Shute, S. J., & Normore, L. F. (1990). Toward the development of semantically-based search systems. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 155-160. doi: 10.7152/acro.v1i1.12475

Smith, P. J., Shute, S., Chignell, M., and Krawczak, D. (1989) Bibliographic information retrieval: Developing semantically-based search systems. *Advances in Man-Machine Systems Research*, *5*, 93-152.

Smith, P. J., Shute, S., Galdes, D., and Chignell, M. (1989) Knowledge-based search tactics for an intelligent intermediary system. *ACM Transactions on Information Systems*, *7*, 246-270.

ISSN: 2324-9773

Smith, P. J., Denning, R., Shute, S. J., & Normore, L. F. (1990). Toward the development of semantically-based search systems. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 155-160.
doi: 10.7152/acro.v1i1.12475

ISSN: 2324-9773