Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval

Bruce R. Schatz
schatz@uiuc.edu

Eric H. Johnson
ejohnson@uiuc.edu

Pauline A. Cochrane
pcochran@uiuc.edu

Digital Library Initiative
Grainger Engineering Library Information Center
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Hsinchun Chen
hchen@bpa.arizona.edu
Department of Management Information Systems
University of Arizona, Tucson

Abstract

The basic problem in information retrieval is that large-scale searches can only match terms specified by the user to terms appearing in documents in the digital library collection. Intermediate sources that support term suggestion can thus enhance retrieval by providing alternative search terms for the user. Term suggestion increases the recall, while interaction enables the user to attempt to not decrease the precision.

We are building a prototype user interface that will become the Web interface for the University of Illinois Digital Library Initiative (DLI) testbed. It supports the principle of multiple views, where different kinds of term suggestors can be used to complement search and each other. This paper discusses its operation with two complementary term suggestors, subject thesauri and co-occurrence lists, and compares their utility. Thesauri are generated by human indexers and place selected terms in a subject hierarchy. Co-occurrence lists are generated by computer and place all terms in frequency order of occurrence together. This paper concludes with a discussion of how multiple views can help provide good quality Search for the Net.

This is a paper about the design of a retrieval system prototype that allows users to simultaneously combine terms offered by different suggestion techniques, not about comparing the merits of each in a systematic and controlled way. It offers no experimental results.

Introduction to search terms

Effective information retrieval on an on-line document collection closely resembles the problem of effectively searching a library catalog by subject. As opposed to a known-item search, where you know what you want from the start and can provide precise title and/or author information, at the start of a subject search you only know that you want documents "about" something. The set of documents you come away with depends on the set of words you provide to the retrieval system and the ways in which it allows you to apply those words to the database. And even when you have a set of documents that appear relevant to your problem, you can never be sure that there are not more documents in the collection that you might find useful. This illustrates the completeness problem inherent in all information retrieval systems.

To attempt greater completeness of a set of retrieved documents, you might combine into one larger set the results of several different searches, each with a different search term. But here the problem of precision arises: as you use more search terms to retrieve a set of documents (assuming a Boolean "or" or set union between each term used), the proportion of documents in that set that you would consider relevant to your problem tends to decrease. Even a retrieval set based on only one search term may contain a lot of irrelevant documents while still excluding many of the relevant documents in the collection. Doing effective information retrieval, then, largely depends on picking search terms that, in your own judgment, yield retrieval sets that contain a high proportion of relevant documents while excluding few, if any, of the other relevant documents in the collection. In short, you need to specify search terms that retrieve relevant documents with completeness and precision.

Picking the "right" search terms for your problem depends on how well you know the vocabulary used in the documents you want to retrieve. Therefore, you can typically get useful results when searching a collection of documents within your own field of expertise, but outside of that you will not have as much success even if you know the desired concepts, because you will not always know the correct terms to use for your search. Despite user knowledge that several terms within a particular domain may have the same meaning, known retrieval technology can only match terms provided by the searcher to terms literally occurring in documents or indexing records in the collection. While techniques like word stemming can improve retrieval somewhat, retrieval based on synonyms and other latent content of documents requires access to auxiliary search term databases outside of the actual document collection.

DL'96, Bethesda MD USA
© 1996 ACM 0-89791-830-4/96/03..$3.50

The use of term suggestion for information retrieval

Information specialists have long worked to bring the vocabularies used by searchers closer to those of the collections they maintain. Conventional library catalogs typically have the benefit of human indexers who assign "aboutness" to documents in the form of subject terms assigned to their bibliographic records. These come from collections of preferred subject terms, called subject thesauri, provided to indexers and searchers alike. The Library of Congress Subject Headings (LCSH) is probably the most widely known example, even though by many criteria it is not a particularly good one [6].

Indexing organizations, such as IEEE or the National Library of Medicine, that concentrate on specialized areas of knowledge tend to produce detailed subject thesauri that present terms in highly organized ways that reflect how subject experts in those fields understand those terms. Subject thesauri also provide synonym control, which reduces the number of different phrases used for a subject search to those used by the indexers of the document collection. The idea is to collapse a set of semantically equivalent terms into one preferred term that you can then use to actually retrieve bibliographic records. Otherwise, you may have the right concept in mind, which in principle should retrieve documents with sufficient completeness and precision, but in practice does not because the authors and indexers used a different term for that concept.

Besides providing vocabulary control for retrieval based on subject headings, thesauri also provide tracings between preferred terms that can suggest broader, narrower, and non-hierarchically related alternatives to the initial search term. Thesauri thus provide a dual function: they help you avoid search terms not used by indexers while they suggest other search terms which have distinct and precise meanings within a number of conceptual schema.

The idea of a thesaurus "suggesting" terms to the searcher is just that: the onus of selecting suitable terms for effective retrieval still rests on the searcher, but the thesaurus reveals much about the indexing of the collection and thus makes the searcher's job much easier. Effective suggesting of terms can come from other mechanisms as well: browseable keyword and keyword-in-context lists, classification schemes, co-occurrence lists, and even bibliographic records with multiple subject term assignments. All of these mechanisms give you external structural cues when searching document collections. But they can only suggest terms to use; you must decide for yourself whether to use them or not.

Each of the term suggestion mechanisms listed above present to the searcher, in their own unique ways, the content of the document collection. Thesauri are constructed over time and change as the collection grows and the terminology of the fields it covers changes. Classification schema (e.g. the Dewey Decimal System) evolve in the same way, but have a more rigid structure in that they try to lay out all terms within a single grand hierarchical sequence. Bibliographic records cluster both thesaurus terms and classification terms around a single document to describe what it is about. All three of these are the result of intellectual effort by professional indexers, and provide indispensable term suggestion mechanisms for locating documents with the same kinds of "aboutness."

Professional indexers only include a term in a thesaurus or a classification scheme if it occurs in the literature that they index. Even then, it must endure a sort of canonization ritual, where it lives as a free text identifier (rather like a blessed free-text term) for a time. If it demonstrates enough usefulness there, and can also fill a gap in meaning in the present version of the thesaurus or classification scheme, only then will the lexicographer add it in the appropriate place. This tendency towards conservatism in thesaurus construction keeps the structure of thesauri stable and the terms within them viable over extended periods.

Co-occurrence lists, in contrast, are the result of intensive statistical calculations on how terms in documents in the collection occur together. The co-occurrence lists for a document collection are selected from a matrix containing the frequency of all pairs of terms occurring within, for example, the same sentence. Given a term, the list of all terms co-occurring with it can then be displayed in frequency order for use in interactive term suggestion. See [3] for a description of algorithms for term co-occurrence analysis. Currently, supercomputers are required to do the necessary computations to create such lists for large collections in a reasonable amount of time.

Different views of the same collection

Each of these term suggestion mechanisms is useful in its own way. A subject thesaurus presents "meaning," which terms are conceptually related to which, while co-occurrence lists present "context," which terms appear in context with which. They are both useful but for different purposes -- the thesaurus for precision, since the hierarchy is "correct," and the co-occurrence lists for recall since they show many more closely "associated" terms. Thus the thesaurus reflects "real" semantics at a gross level while the co-occurrence reflects "real" documents at a finer level (since all the words from the documents are included giving recent coverage but without human discretion as to their meaning).

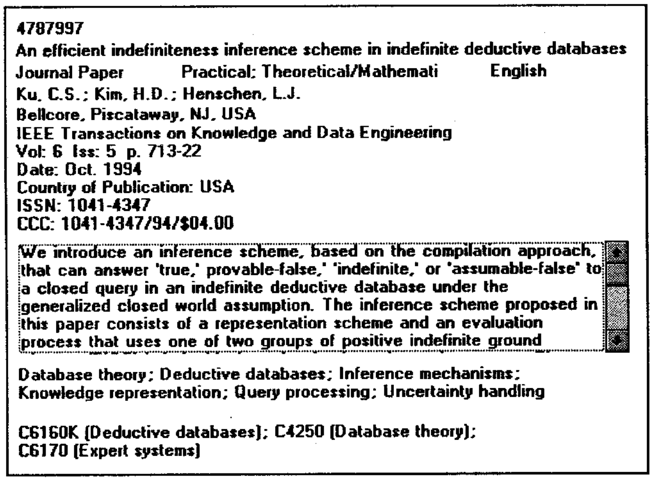
By providing easy access to these various term suggestion mechanisms, we hope to encourage searchers to use them before attempting to access bibliographic records. This would reverse the current state of bibliographic as well as full-text retrieval, in which thesauri and other means of term suggestion (assuming they are even available) are typically accessed only after an initial bibliographic query yields either too few or too many hits.

In this paper we compare the use of two of the term-suggestion mechanisms described above, subject thesauri and co-occurrence lists, and in doing so show how each complements the other. For our research, we have been using the INSPEC™ Thesaurus as a sample thesaurus, and the "concept space" generated from 400,000 INSPEC indexing and abstracting records as a sample database of co-occurrence lists. These reflect roughly the same document collection indexed by humans and by computer respectively. (INSPEC is the indexing and abstracting service covering most of the research literature in physics, electrical engineering, and computer science. It is maintained by the Institution of Electrical Engineers, the British equivalent of the IEEE.)

Our research thus far has involved constructing a prototype system to provide interactive term suggestion to searchers of digital libraries, to be incorporated into the University of Illinois DLI testbed. Usability studies are planned during the next six to eight months to test the effectiveness of this system. Here we describe the possibilities that thesaurus and co-occurrence list browsing in particular and simultaneous use of multiple auxiliary views of a collection in general can offer to users of information retrieval systems. The examples which follow are taken from sessions with our prototype of such a system.

Documents and bibliographic records

Different term suggestion mechanisms present to the user very different views of a bibliographic collection. Their creation, as well as their use, are dictated by the very different requirements and expectations of various types of indexing, or, in terms of retrieval, by the different ways in which controlled vocabularies and natural language are used.



4787997
An efficient indefiniteness inference scheme in indefinite deductive databases
Journal Paper Practical: Theoretical/Mathemati English
Ku, C.S.; Kim, H.D.; Henschen, L.J.
Bellcore, Piscataway, NJ, USA
IEEE Transactions on Knowledge and Data Engineering
Vol: 6 Iss: 5 p. 713-22
Date: Oct. 1994
Country of Publication: USA
ISSN: 1041-4347
CCC: 1041-4347/94/$04.00

We introduce an inference scheme, based on the compilation approach, that can answer 'true,' provable-false,' 'indefinite,' or 'assumable-false' to a closed query in an indefinite deductive database under the generalized closed world assumption. The inference scheme proposed in this paper consists of a representation scheme and an evaluation process that uses one of two groups of positive indefinite ground

Database theory; Deductive databases; Inference mechanisms; Knowledge representation; Query processing; Uncertainty handling

C6160K (Deductive databases); C4250 (Database theory); C6170 (Expert systems)

Figure 1. Sample INSPEC bibliographic record.

Figure 1 illustrates a typical INSPEC bibliographic record as displayed by our prototype interface. Document surrogates such as this are what we search for when doing retrieval, and what we would like thesauri and co-occurrence lists to help us find. Like the actual document it represents, the bibliographic record contains full title and author information, as well as the abstract as it appears in the document. Below the abstract in figure 1 (in the scrollable area) appear the indexing terms taken from the INSPEC Thesaurus (e.g. Database theory, Deductive databases, Inference mechanisms, etc.). At the bottom of the bibliographic record appear the classification codes and captions, which together constitute another term suggestion mechanism we plan to use but do not discuss in this paper.

The value added to a bibliographic record by having human indexers assign indexing terms to it can be seen in figure 1. The article is about deductive databases, and the term "deductive databases" appears in the title, subject terms (from the thesaurus), and the classification terms. The term "deductive database", a stem of "deductive databases", appears in the abstract. We could thus retrieve this record with the search term "deductive databases" with a title or subject search, or with a text search if the system we use supports word stemming. But such strong concurrence between title, text, and subject terms is rare. Assuming that the indexer did a good job of determining the aboutness of the article, it is also about database theory (an admittedly broad term), inference mechanisms, knowledge representation, query processing, and uncertainty handling. The list of classification terms adds expert systems as well. None of the phrases "database theory", "knowledge representation", "uncertainty handling", or "expert systems" actually occur in the title or abstract, so a title or text search using any of those terms would not retrieve this record. Only a subject index search would. Similarly, "inference mechanisms" does not occur in the abstract or title, but "inference scheme" does; searching on the single keyword "inference" would retrieve this record, but would reduce the precision of the result set. The same is true for "query processing": "query evaluation" occurs in the abstract, but as with "inference" searching on the single keyword "query" would reduce the precision of the result set.

By using the controlled vocabulary supplied by a thesaurus, search-broadening techniques like word stemming and proximity searching are not necessary to retrieve records that cover the same concept, because indexers use the same term for that concept throughout the bibliographic database. This prevents records with other terms that represent different concepts, but match a stemmed or a word proximity query based on the desired term, from being retrieved, thus helping preserve the precision of the retrieved set of records.

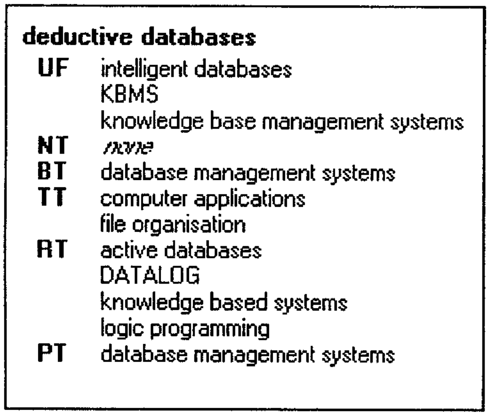
A thesaurus, however, does not perfectly cover its subject domains, nor does it control all synonyms for the terms within them. "Inference scheme" and "query evaluation" from the sample bibliographic record in figure 1 are two such terms. This is due in part to the lag time involved in the canonization process described above, as well as the inability of even the most thorough lexicographer to catch every synonym for a concept in the literature. This is where computer-generated term suggestion mechanisms are most helpful. What they lack in semantic substance and conceptual precision they make up for in completeness and currentness.

Differences in both form and use between subject thesauri and co-occurrence lists are best illustrated with an example of what each might present during the term selection process. As we have already suggested, these should offer complementary yet comparable ways of retrieving records from a database of bibliographic records or full-text documents.

Subject thesaurus display

Figure 2 shows the INSPEC Thesaurus record for the preferred term "deductive databases", one of the terms used to index the sample bibliographic record in figure 1. Notice that it lists a number of Use For references, indicated by the "UF" tracing label; these correspond to references in the thesaurus from the terms "intelligent databases", "KBMS", and "knowledge base management systems" to the preferred term. Terms indicated by the NT tracing label are Narrower Terms (in this case none) of the term; by BT, Broader Terms; by TT, Top Terms; by RT, Related Terms (considered associated but not in a discernibly hierarchical way); and by PT, Prior Terms (like UFs, but used at some time previously to index items now indexed in some cases with the current preferred term).

The thesaurus record shown in figure 2 is essentially how it appears in the current printed edition of the INSPEC Thesaurus. It provides links to other terms, but the overall scheme into which it fits is difficult to discern. By flipping pages to other entries you can reconstruct the hierarchy and perhaps find RT tracings that interest you, but this is tedious and time-consuming.



deductive databases

UF	intelligent databases KBMS knowledge base management systems
NT	*none*
BT	database management systems
TT	computer applications file organisation
RT	active databases DATALOG knowledge based systems logic programming
PT	database management systems

Figure 2. INSPEC Thesaurus record for the preferred term "deductive databases".

The prototype thesaurus browser we have developed provides a visual representation of a thesaurus by reconstructing the disembodied conceptual schema scattered among the thousands of entries in a typical thesaurus. The lefthand side of figure 3 shows a partial view of the INSPEC thesaurus entry for "deductive databases" as displayed by our prototype thesaurus browser. (Figure 5 shows the thesaurus browser display as it might appear along with other windows on the computer screen.)

To construct a visual display for a typical entry, the thesaurus browser must use data from other thesaurus entries as well. Specifically, the thesaurus display for "deductive databases" on the lefthand side of figure 3 requires the browser to use NT tracings from the entry for "database management systems" as well as the entry for "information systems" and a number of other entries not shown.

To briefly explain the layout and function of the thesaurus browser display, scope notes and related term tracings (RTs) for the current term are taken directly from the thesaurus record and shown in the display under the phrase "Terms related to...". The hierarchy (compiled from BT, NT, and TT tracings) in which the current term occurs is the principal section of the display. A little triangle next to an entry in the hierarchy (here called an "expansion triangle") indicates that it has narrower terms, and whether the triangle is upright (pointing to the right) or tipped over (pointing down) indicates whether the hierarchy under the term is collapsed or expanded, respectively.

The thesaurus browser is completely hypertextual. You can click on any term you see on the display to see the entry for that term (which then becomes the current term). When the browser displays the entry for a thesaurus term, it automatically expands appropriate parts of the hierarchy and displays the term in boldface to clearly show you where it occurs, while leaving other parts of the hierarchy unexpanded. This yields a "fisheye" view of the term in the hierarchy, with the parts of the hierarchy near it expanded and the parts away from it left unexpanded.

See [5] for a more detailed description of how the thesaurus browser works and how it relates to the structure of the particular thesaurus it displays.

Co-occurrence list display

The right-hand side of figure 3 shows the co-occurrence list for the term "deductive databases" in the concept space generated from the INSPEC indexing and abstracting records mentioned above. The terms listed at right appear in decreasing order of the weight of their co-occurrence with the term "deductive databases": "database theory", "logic programming", and "query languages" are the three highest weighted co-occurring terms. The further down the co-occurrence list a term appears, the less often it occurs along with "deductive databases" in the INSPEC database.

Like the thesaurus browser, the co-occurrence list browser is hypertextual: clicking on a term in a co-occurrence list yields the weighted list for that term. Both browsers allow you to navigate their respective "spaces" by following links in this way.

See [2] for a more detailed description of co-occurrence lists and an explanation of the algorithms used to generate them.

Comparing subject thesauri with co-occurrence lists

Unlike a thesaurus, there is no structure to the relationships in a co-occurrence list; only the weights of the links between the co-occurring terms. The statistical procedures employed to generate co-occurrence lists cannot discern a term's meaning and scope of application as humans can, and thus cannot discern whether a term is "broader" or "narrower" than another and assign a BT or NT relationship, respectively. Considering only the kinds of relations expressed in a subject thesaurus, the best that a co-occurrence list can manage is something like an RT (related term), where there is no discernible hierarchical relationship between the terms, though they are still considered to be associated in one way or another. Nor do co-occurrence lists make any attempt at synonym control or other kinds of vocabulary restriction.

These appear as shortcomings only if you try to use co-occurrence lists the same way you would use a thesaurus. A thesaurus gives precision to the meanings of the terms you use for retrieval, while a co-occurrence list aids in recall by revealing the context in which terms are used in the collection, be they thesaurus terms or other "uncontrolled" terms. In this way they offer terms to use in searches of fields of bibliographic records containing unrestricted vocabularies, such as the title or abstract, or even to full-text searching in systems that offer it. They may also aid in searching the thesaurus itself by suggesting terms not visible in the currently displayed hierarchy.

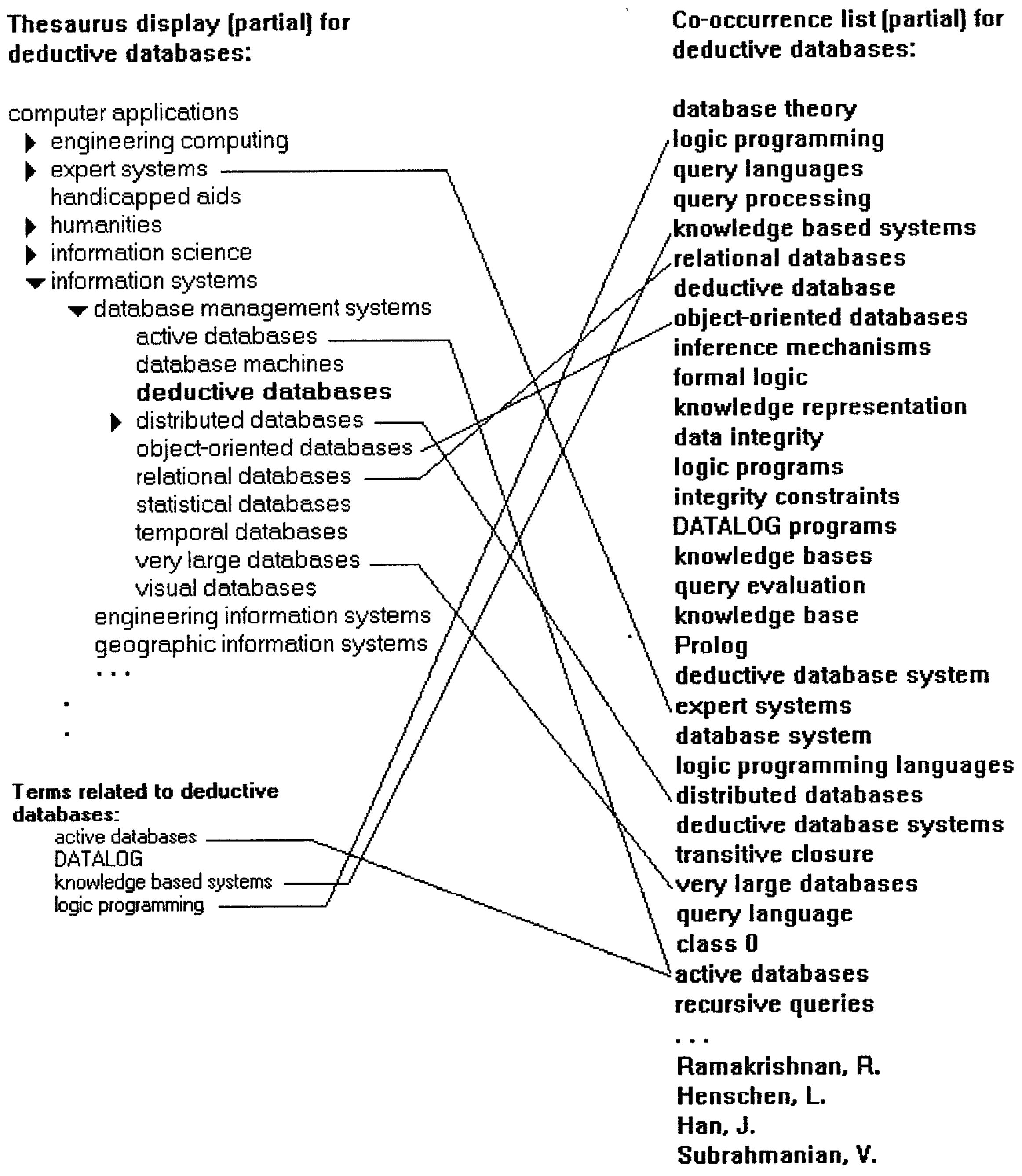
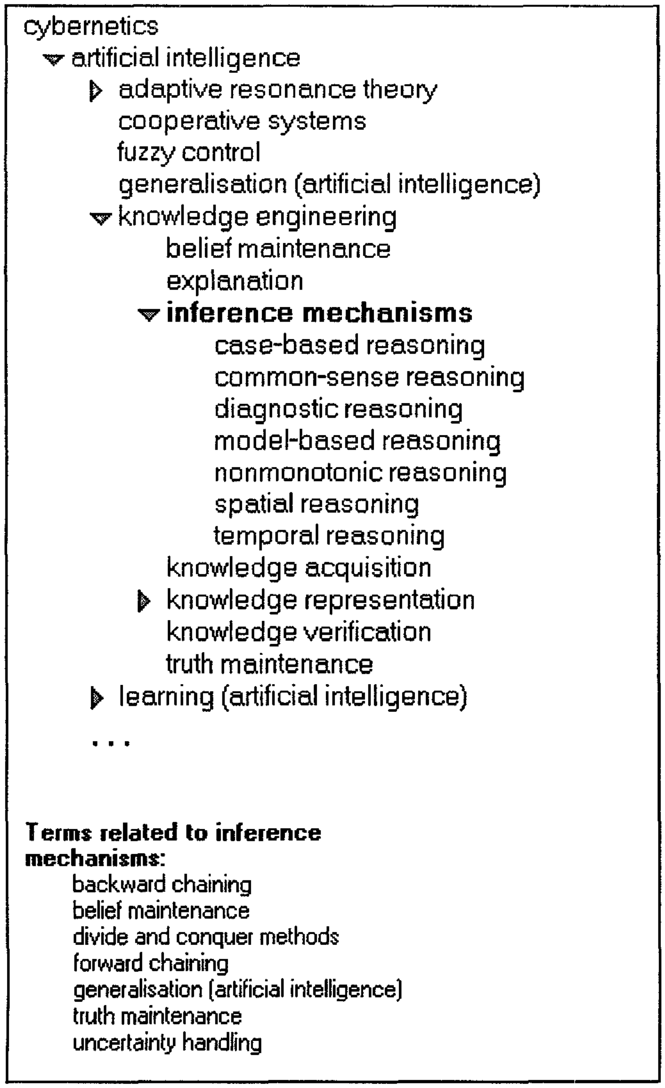


Figure 3. Comparative views of thesaurus and co-occurrence list content for the term "deductive databases" (arrangement of terms has been altered somewhat to facilitate comparison between lists). Lines conneat terms that occur in both.

Earlier we suggested that thesaurus terms appearing in a co-occurrence list for a given term were akin to related terms rather than broader or narrower terms, and this example illustrates such a case. All RTs (terms listed at lower left under the heading "Terms related to deductive databases") are in the co-occurrence list (arguably, the term "DATALOG" occurs in the co-occurrence list as part of "DATALOG programs".)

Recall that in the sample bibliographic record in figure 1, "database theory" and "inference mechanisms" appear with "deductive databases" as subject terms used to index the record. The appearance of the former two terms in the co-occurrence list for "deductive databases" is in part due to their co-occurrence in this indexing record.

The suggestion of other terms, in taking you to different parts of the thesaurus, can suggest yet more search terms. Figure 4 illustrates the result of entering "inference mechanisms", suggested by the co-occurrence list, into the thesaurus. You can play the thesaurus and the co-occurrence lists together this way to get as many search terms related to your problem as you want. This is explored in the next section.



```
cybernetics
  ▽ artificial intelligence
      ▷ adaptive resonance theory
        cooperative systems
        fuzzy control
        generalisation (artificial intelligence)
      ▽ knowledge engineering
            belief maintenance
            explanation
          ▽ inference mechanisms
                case-based reasoning
                common-sense reasoning
                diagnostic reasoning
                model-based reasoning
                nonmonotonic reasoning
                spatial reasoning
                temporal reasoning
            knowledge acquisition
          ▷ knowledge representation
            knowledge verification
            truth maintenance
      ▷ learning (artificial intelligence)
      . . .
```

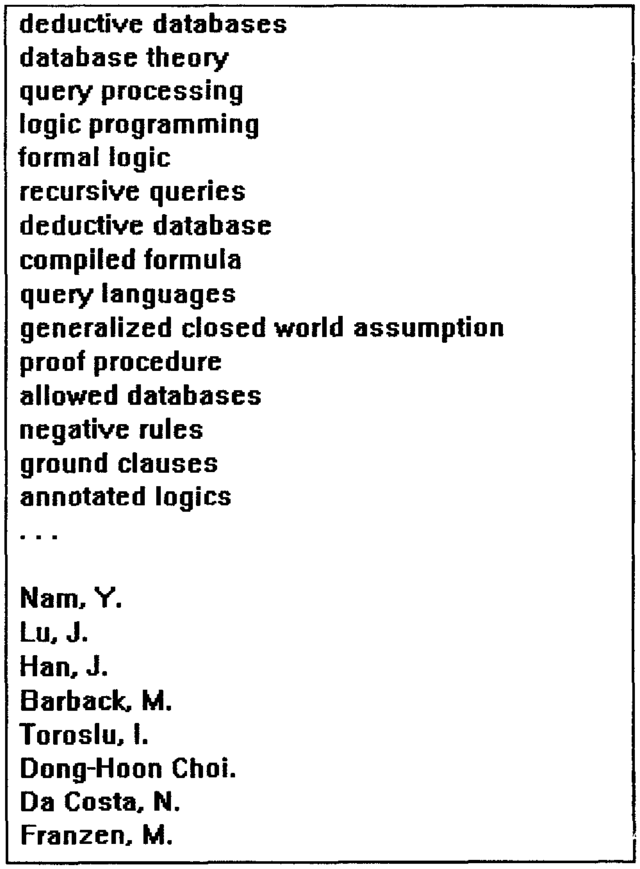
Terms related to inference mechanisms:

backward chaining
belief maintenance
divide and conquer methods
forward chaining
generalisation (artificial intelligence)
truth maintenance
uncertainty handling

Figure 4. Partial thesaurus browser display for the term "inference mechanisms".

Another term near the top of the co-occurrence list for "deductive databases" is "query evaluation", which appears in the abstract of the sample bibliographic record in figure 1 (though it is not visible because it is in the bottom part of the abstract, scrolled out of view). "Query evaluation" is not a thesaurus term, probably because "query processing" covers the same concept, even though there is no USE tracing in the thesaurus from the former to the latter. "Query evaluation" is therefore a free-text term suggested by the co-occurrence list which can be used in a free-text search to retrieve the bibliographic record.

This in turn suggests a role for co-occurrence lists in thesaurus construction, in that they can complement the work of the human lexicographer mentioned earlier. In the "query evaluation / query processing" example above, the co-occurrence list for the preferred term "query processing", by suggesting to the lexicographer a USE tracing from "query evaluation", could help improve the synonym control provided by the thesaurus.



deductive databases
database theory
query processing
logic programming
formal logic
recursive queries
deductive database
compiled formula
query languages
generalized closed world assumption
proof procedure
allowed databases
negative rules
ground clauses
annotated logics
. . .

Nam, Y.
Lu, J.
Han, J.
Barback, M.
Toroslu, I.
Dong-Hoon Choi.
Da Costa, N.
Franzen, M.

Figure 5. Partial co-occurrence list for the author Henschen, L.

Yet another bonus of a co-occurrence list is that, because it lists the context of all text items in a collection, it also lists the context of author names, as well as the author names that co-occur with conceptual terms, as shown at the bottom of figure 3. This is outside of the scope of a thesaurus, which at best offers named phenomena and devices, e.g. "de Broglie waves", "van de Graaff accelerators". Names include more than authors (e.g. programming languages such as "Prolog") and other proper nouns. Human indexers, in attempting to include only concepts in subject thesauri, leave out proper nouns, which are very useful search terms indeed! The computer programs that generate co-occurrence lists include personal names and other proper nouns because they can recognize strings but not the human-understood meanings of them.

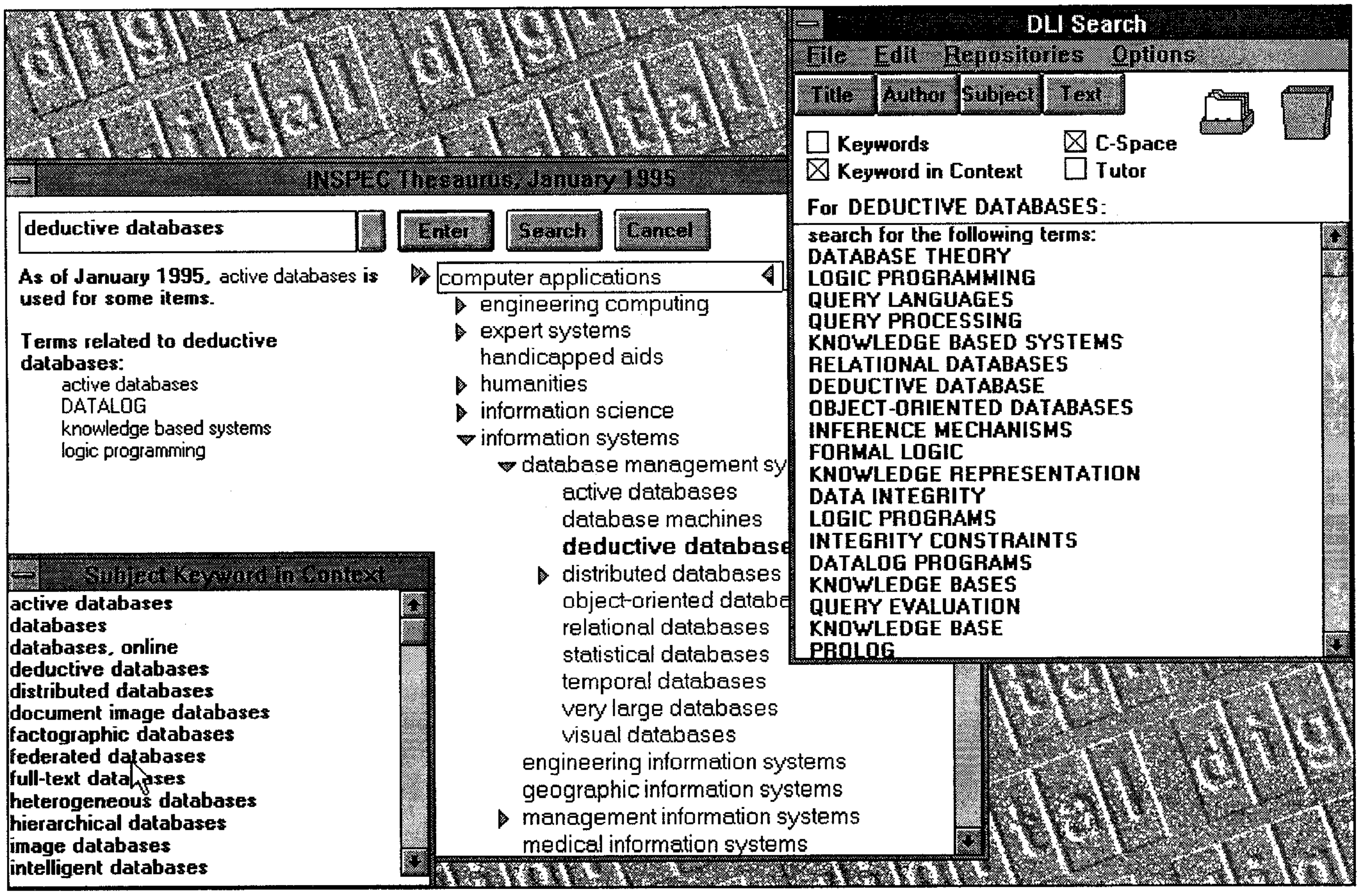
Having an author name in the conceptual area you are searching gives you a powerful search term, with which you can quickly gather a set of fruitful bibliographic records, from which in turn you can gather additional subject terms and keywords, as well as other authors. Assuming that the author has a somewhat narrow research field, which is typically the case in academic research, the retrieved record set will have good precision as well.

One of the authors in the co-occurrence list in figure 3, "Henschen, L.", is one of the authors of the article whose bibliographic record is in figure 1. Figure 5 illustrates the result of entering this author into the co-occurrence list display. Compare this with figure 3. In this way, and in ways demonstrated above, the semantic looseness and lack of structure of co-occurrence lists provides a powerful way to move among and between more structured information spaces.

Using multiple simultaneous views

Much can be inferred from the previous example about the general usefulness of having different kinds of views of a collection, as well as the usefulness of views that offer different degrees of semantic substance and structure. Recall that the manual thesaurus browser arranges terms according to human conceptual relationships, while co-occurrence lists show terms (whether in the thesaurus or not) that are contextually related to a given term. They are both useful but for different purposes -- the thesaurus for precision since the hierarchy is "correct" and the co-occurrence lists for recall since they show many more terms.

When these and other term suggestion mechanisms are combined in our prototype multiple view interface, they allow you to quickly drag and drop terms from one view to another. [4] The techniques for searching offered by each view are thus available simultaneously. It is well known in the information science literature that there are different kinds of search needs and that user behavior is best facilitated by providing different search interfaces tuned to each particular need. Figure 6 illustrates part of the session from which we gathered the previous five figures.


DLI Search
File Edit Repositories Options
Title Author Subject Text
INSPEC Thesaurus, January 1995
Enter Search Cancel
Subject Keyword in Context

☐ Keywords ☒ C-Space
☒ Keyword in Context ☐ Tutor

For DEDUCTIVE DATABASES:

search for the following terms:
DATABASE THEORY
LOGIC PROGRAMMING
QUERY LANGUAGES
QUERY PROCESSING
KNOWLEDGE BASED SYSTEMS
RELATIONAL DATABASES
DEDUCTIVE DATABASE
OBJECT-ORIENTED DATABASES
INFERENCE MECHANISMS
FORMAL LOGIC
KNOWLEDGE REPRESENTATION
DATA INTEGRITY
LOGIC PROGRAMS
INTEGRITY CONSTRAINTS
DATALOG PROGRAMS
KNOWLEDGE BASES
QUERY EVALUATION
KNOWLEDGE BASE
PROLOG

deductive databases

As of January 1995, active databases is used for some items.

Terms related to deductive databases:
active databases
DATALOG
knowledge based systems
logic programming

computer applications
- engineering computing
- expert systems
- handicapped aids
- humanities
- information science
- information systems
  - database management sy
    - active databases
    - database machines
    - **deductive database**
    - distributed databases
    - object-oriented databa
    - relational databases
    - statistical databases
    - temporal databases
    - very large databases
    - visual databases
  - engineering information systems
  - geographic information systems
  - management information systems
  - medical information systems

active databases
databases
databases, online
deductive databases
distributed databases
document image databases
factographic databases
federated databases
full-text databases
heterogeneous databases
hierarchical databases
image databases
intelligent databases

Figure 6. Session screen showing actual appearance of thesaurus and co-occurrence list displays. The keyword in context display is also shown, in the lower lefthand corner.

Multiple views are the user interface principle around which we are building the prototype described in this paper. In the coming year we are extending it to become the "Web interface" for the Digital Library Initiative (DLI) project at the University of Illinois at Urbana-Champaign. This large-scale digital library testbed is building a collection of SGML documents, consisting of articles from magazines and journals in engineering and science obtained in a direct pipeline from major technical publishers. The multiple view interface will support easy combination of term suggestion from different sources followed by full-text search of the document collection. As the SGML collection primarily covers computer science, electrical engineering, and physics, we are using the INSPEC subject thesaurus and parts of the Dewey Decimal Classification supplemented with co-occurrence lists from the bibliographic areas being covered. Since the plans in the coming three years are to build a testbed with 100,000 documents from many publishers and 100,000 users across the Big Ten universities, the extensive sociological evaluation will provide a large-scale test of the utility of the multiple view principle for information retrieval. More details on the DLI project are contained in [7].

The true utility of multiple views will only become apparent when we have many sources to combine seamlessly. The DLI testbed efforts will experiment with 2 or 3 sources of 2 or 3 kinds. An on-going experiment in the DLI research efforts (the longer-term portions of the project) will greatly extend this to an experiment with hundreds of term suggestion sources rather than just a few as we have now. Over the next few months, we plan to generate co-occurrence lists for all of engineering. Some 3 million abstracts from Compendex (Engineering Index), which has broad coverage across all of engineering, will be used as materials for the generation of fine-grained co-occurrence lists. These materials can be thought of as 15 broad areas with 200,000 abstracts each (roughly the size of the INSPEC collection which covers only 2 areas). However, we plan to divide these along Compendex class code lines to get much smaller areas covered by subjects like "bridges" rather than the much larger areas covered by all of civil engineering, for instance. This will yield hundreds of co-occurrence lists relevant for term suggestion across our user population (faculty and students in engineering at the University of Illinois). Since the INSPEC experiment showed that computing a list for an area required roughly a day (24 hours) of supercomputer time, our experiment is possible only because the NCSA (National Center for Supercomputing Applications) is granting us special time on their newest computer (Convex Exemplar) during its testing phase.

During the coming year we will rewrite the DLI Web interface from its current implementation in Microsoft Visual Basic to a production version implemented in Java^{TM}. Java is a new protected execution environment being bundled into Web browsers such as Netscape^{TM} that enables dynamic loading of programs and data across the Internet. The plan is for our Web interface to have generic support for term suggestion in general and for subject thesauri, co-occurrence lists, and full-text search in particular. When organized collections become more common on the Web, this leads to the possibility of dynamically loading the term suggestors for all the collections desired by the user on a demand basis. Transparent Multiple Views will be a necessity when users casually perform search sessions spanning hundreds of thousands of fine-grained subject domains. The experiments in the Illinois DLI project on interactive term suggestion will give the first taste and develop the first good technology for Search in the Net.

Acknowledgments

We wish to thank Jim Ashling of IEE England and Michele Day, Manager of the INSPEC US Office, for making the INSPEC Thesaurus available to us in machine-readable form, and for permission to use the INSPEC A&I records for generating term co-occurrence lists.

This project is funded by NSF/ARPA/NASA Digital Library Initiative DLI grant to University of Illinois IRI-94-11318COOP.

References

[1] R. Allen (1994). Navigating and Searching in Hierarchical Digital Library Catalogs, Digital Libraries '94 Proceedings, College Station, TX June 19-21, 1994, pp. 95-100.

[2] H. Chen, B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, C. Lin (1995) A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project, submitted to IEEE Trans Pattern Analysis and Machine Intelligence, Special Issue on Digital Libraries: Representation and Retrieval, 15pp.

[3] H. Chen, B. Schatz, T. Yim, D. Fye (1995) Automatic Thesaurus Generation for an Electronic Community System, Journal American Society Information Science 46 (3): 175-193, April 1995.

[4] E. Johnson (1995) Extending an Interactive Thesaurus by Dragging, ACM SIGLINK Newsletter, pp. 16-17 (September).

[5] E. Johnson, P. Cochrane (1995) A Hypertextual Interface for a Searcher's Thesaurus, Digital Libraries '95 Proceedings, Austin, TX June 11-13, 1995, pp. 77-86.

[6] M. Kirtland, P. Cochrane (1981) Critical Views of LCSH: A Bibliographic Essay. ERIC Document ED 208900.

[7] B. Schatz, B. Mischo, T. Cole, J. Hardin, L. Jackson, A. Bishop, L. Star, P. Cochrane, H. Chen (1996) Digital Library Infrastructure for a University Engineering Community: Towards Search in the Net via Structure and Semantics, submitted to IEEE Computer, Special Issue on Large-Scale Digital Libraries, 12pp.