Alleviating Search Uncertainty through Concept Associations: Automatic Indexing, Co-Occurrence Analysis, and Parallel Computing

Hsinchun Chen
MIS Department, Karl Eller Graduate School of Management, University of Arizona, McClelland Hall 430Z, Tucson, AZ 85721. E-mail: hchen@bpa.arizona.edu

Joanne Martinez
University Libraries, University of Arizona, Science-Engineering Library, Room 209, P.O. Box 210054, Tucson, AZ 85721. E-mail: martinez@bird.library.arizona.edu

Amy Kirchhoff
Princeton University, 87 Prospect Avenue, Princeton, NJ 08544. E-mail: amykir@princeton.edu

Tobun D. Ng
MIS Department, Karl Eller Graduate School of Management, University of Arizona, McClelland Hall 430, Tucson, AZ 85721. E-mail: tng@bpa.arizona.edu

Bruce R. Schatz
Graduate School of Library and Information Science, NCSA, Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews Ave., Urbana, IL 61801. E-mail: schatz@csl.ncsa.uiuc.edu

In this article, we report research on an algorithmic approach to alleviating search uncertainty in a large information space. Grounded on object filtering, automatic indexing, and co-occurrence analysis, we performed a large-scale experiment using a parallel supercomputer (SGI Power Challenge) to analyze 400,000+ abstracts in an INSPEC computer engineering collection. Two system-generated thesauri, one based on a combined object filtering and automatic indexing method, and the other based on automatic indexing only, were compared with the human-generated INSPEC subject thesaurus. Our user evaluation revealed that the system-generated thesauri were better than the INSPEC thesaurus in *concept recall,* but in *concept precision* the 3 thesauri were comparable. Our analysis also revealed that the terms suggested by the 3 thesauri were complementary and could be used to significantly increase "variety" in search terms and thereby reduce search uncertainty.

1. Introduction

Large electronic information storage and retrieval systems and databases such as online catalogs, online bibliographic databases, legal and finance databases, WWW servers, and video databases are changing the way we gather, process, and retrieve information. These systems provide a wide variety of information and services, ranging from daily updates of foreign and national news, movie reviews and clips, law cases, and financial data on companies to journal articles, books, trademarks, and statistics. However, gaining access to such information is often difficult. This is due, in large part, to the indeterminism involved in the process by which information is indexed, and to the latitude searchers have in expressing a query.

2. Using Thesauri to Alleviate Search Uncertainty: Literature Review

2.1. Indexing and Search Uncertainty

The process of indexing is partly indeterminate. Evidence suggests that different indexers, well trained in an indexing scheme, might assign index terms for a given document differently. It has also been observed that an indexer might use different terms for the same document at different times (Jacoby & Slamecka, 1962; Stevens, 1965).

Search uncertainty refers to the latitude searchers have in choosing search terms. An even higher degree of uncertainty with regard to search terms has been observed (compared to indexing uncertainty). Searchers tend to use different search terms for the same information. Studies have revealed that, on average, the probability of any two people using the same term to describe an object is less than 20% (Furnas, Landauer, Gomez, & Dumais, 1987; Gomez & Lochbaum, 1984; Good, Whiteside, Wixon, & Jones, 1984). This limits the success of various design methodologies for controlled vocabulary-driven interaction (Chen, 1994; Furnas et al., 1987).

Bates (1986) argued that for a successful match, the searcher must, to some extent, generate as much ‘‘variety’’ in the search as has been produced by the indexers in their indexing. The variety produced by an indexer can also be viewed as *redundancy* in the sense that there is partial overlapping of the classifications applied to a document. To increase the chances of a successful match, a number of indexes for each document should be available, and the redundancy (generated by the indexer) associated with each document should be preserved. However, in practice, catalog systems discourage redundancy for the following reasons (Bates, 1986; Chan, 1986):

- *Whole document indexing:* A cataloger working according to the Library of Congress Subject Headings or some other indexing schemes is trained to index the whole document, not parts or concepts within it.
- *Specific entry:* Each document is to be entered under a category (heading) which is specific to the content, neither broader or narrower in scope than the scope of the document’s contents.
- *Limited cross-reference structure:* Cross references are frequently an afterthought to ‘‘augment’’ the basic catalog organization (Bates, 1977).

In summary, conventions adopted to reduce redundant indexes decrease the likelihood that a searcher will generate the right term for retrieval. Recognizing this problem, reference librarians often rely on extensive thesaurus consultation and searcher query refinement in an attempt to generate the ‘‘variety’’ associated with the search terms and to increase their chances of matching index terms. In our research, we take indexing uncertainty as given. We focus only on improving the search process, assuming that some level of indexing uncertainty will always exist.

2.2. Human-Generated Thesauri and System-Generated Thesauri

Gomez et al. (Furnas et al., 1987; Gomez, Lochbaum, & Landauer, 1990) found in their studies that ‘‘searcher success is markedly improved by greatly increasing the number of names per object.’’ Many research groups have attempted to generate ‘‘variety’’ in search terms by making use of existing thesauri. While these human-generated thesauri are able to provide alternate terms to use in searching, they do not overcome the *knowledge acquisition bottleneck* (Hayes-Roth, Waterman, & Lenat, 1983): The cognitive demand required of humans (indexers or domain experts) to create and maintain thesauri. An alternative approach to creating vocabulary-based search aids is based on *automatic thesaurus generation.* In general, human-generated thesauri are more precise and semantics-rich. System-generated thesauri, on the other hand, are often more comprehensive.

Several large-scale projects have attempted to incorporate human-generated thesauri in the search process. For example, the National Library of Medicine’s Unified Medical Language System (UMLS) project aims to build an intelligent automated system that understands biomedical terms and their interrelationships, and uses this understanding to help users retrieve and organize information from machine-readable sources (Lindberg & Humphreys, 1990; McCray & Hole, 1990). The UMLS includes a Metathesaurus, which consists of biomedical concepts and their relationships as presented in more than 10 different existing vocabularies and thesauri. Chamis (1991) has discussed the issues of thesaurus compatibility and strategies, and systems developed to overcome difficulties in searching multiple incompatible databases. In particular, she has described the effectiveness of the Vocabulary Switching System (VSS), an integrated vocabulary consisting of 12 existing thesauri in four diverse subject areas (business, social sciences, life sciences, physical sciences). Knapp’s BRS/TERM vocabulary database maps natural language synonyms and controlled vocabulary descriptors from seven bibliographic databases in the social and behavioral sciences (Knapp, 1984). The National Technical Information Service (NTIS) database consists of records from databases of numerous government agencies, each of which has its own thesaurus. The NTIS thesaurus represents a merged vocabulary from various micro-thesauri, together with natural language terms, and ‘‘tags’’ indicating the source of each term (Piternick, 1984). In a similar effort, Chaplan (1995) mapped terms from the Laborline Thesaurus to the Library of Congress Subject Headings (LCSH). Development of the Art and Architecture Thesaurus (AAT) began as an attempt to improve upon the LCSH vocabulary by integrating terms from numerous disparate domain-specific thesauri and word lists, and presenting them in a hierarchical structure similar to that of the NLM’s Medical Subject Headings (MeSH). The result is a faceted, hierarchical vocabulary that is compatible with, and appropriate for, libraries primarily centered on LCSH (Petersen, 1983, 1990). Another project undertaken by the Genentech library, based on the methods used by Petersen with the AAT, attempted to rectify inconsistencies between the LCSH and MeSH descriptors in domains related to genetic engineering and molecular biology (Bellamy & Bickham, 1989). Finally, Niehoff and associates at Battelle Columbus Laboratories have developed an integrated vocabulary for the energy domain which represents terms from 11 existing vocabularies (Niehoff, 1976; Niehoff & Kwansy, 1979).

Numerous investigators have developed algorithmic approaches to *automatic thesaurus generation.* Most of these approaches employ techniques that compute coefficients of ''relatedness'' between terms by using statistical co-occurrence algorithms (e.g., cosine, Jaccard, Dice similarity functions) (Chen & Lynch, 1992; Crouch, 1990; Rasmussen, 1992; Salton, 1989). Some algorithms, however, perform cluster analysis to further group terms of similar meanings (Everitt, 1980; Rasmussen, 1992). For example, Crouch and Yang (1992) automatically generated from text keywords thesaurus classes, which can subsequently be used to index documents and queries. Crouch's approach is based on Salton's vector space model and the term discrimination theory. Documents are clustered using the complete link clustering algorithm (agglomerative, hierarchical method). Several research groups have experimented with cross-domain term switching using a combination of human-generated and system-generated thesauri. Chen et al. experimented extensively in generating, integrating, and activating multiple thesauri (some were existing thesauri, others automatically generated, all were in computing-related areas) (Chen, Lynch, Basu, & Ng, 1993; Chen & Ng, 1995). Both Kim and Kim (1990) and Chen et al. (1993) proposed treating thesauri (human-generated and system-generated) as neural networks or semantic networks and applying spreading activation algorithms for term-switching.

3. Generating a Domain-Specific Thesaurus Automatically: Automatic Indexing, Co-Occurrence Analysis, and Parallel Computing

In our previous research, we adopted *object filtering, automatic indexing,* and *co-occurrence analysis* in generating domain-specific thesauri in different domains, e.g., Russian computing (Chen & Lynch, 1992; Chen et al., 1993), business (Chen, Hsu, Orwig, Hoopes, & Nunamaker, 1994), and molecular biology (Chen, Schatz, Yim, & Fye, 1995). We present below a brief overview of these techniques in the context of our recent experiment, which involved a large-scale test collection of 400,000+ computer science and electrical engineering abstracts (1992–1994) acquired from the INSPEC database.[1] Our research goal was to generate a comprehensive computer engineering thesaurus automatically and to compare it with the human-generated INSPEC thesaurus to determine its usefulness in alleviating search uncertainty. Due to the size of the collection, parallel computing was adopted in our experiment. This project is the ''semantic retrieval'' research component of the ongoing NSF/ARPA/NASA funded Illinois digital library project: ''Building the Interspace: Digital Library Infrastructure for a University Engineering Community.''

A. Automatic Indexing and/or Object Filtering

In our previous research, for each online document, we first identified terms that matched with terms in some known vocabularies (which were acquired from different sources), a process referred to as *object filtering.* For most domain-specific databases, there appear always to be some existing lists of subject descriptors (e.g., the subject indexes at the back of textbooks), researchers' names (e.g., author indexes or researchers' directories), and other domain-specific objects (e.g., genes, experimental methods, organizational names, etc.) which exist online or can be obtained through Optical Character Recognition (OCR) scanning. These domain-specific keywords can be used to help identify important concepts in documents automatically. Because the texts remaining after object filtering may still contain many important concepts, an automatic indexing procedure then was followed. In our research, we adopted a general automatic indexing procedure, which includes word identification, stop-wording, and term-phrase formation. The algorithm first identified individual words. Then, a stop-word list was used to remove non-semantic bearing functional (and high-frequency) words such as the, a, on, in, etc. After removing the stop words, term-phrase formation that formulates phrases by combining only adjacent words (2 and 3 words) was performed.

• *A combined object filtering and automatic indexing method*

In our initial experimentation on the 400,000+ INSPEC collection, we adopted a combined object filtering and automatic indexing method. A large object filter list which contained subject descriptors (236,137 descriptors, 4 MBs in size) and author names (334,426 names, 5 MBs) were generated by preprocessing all selected tagged fields of the 400,000+ INSPEC abstracts, i.e., the MARC 650 fields (INSPEC thesaurus terms), the MARC 651 fields (INSPEC indexer-selected terms), and the MARC 100 fields (authors). Following object filtering, a general automatic indexing procedure, as described above, was then adopted to form other new candidate terms. A term occurrence threshold of three was also adopted to remove incidental noise (i.e., a term needed to appear in at least three abstracts to be included in automatic thesaurus generation). Our initial experiment and analysis revealed that, among the 15–25 index terms the system generated for each document, 80% were the results of object filtering and only 20% were derived from automatic indexing. This finding was not surprising but was somewhat disturbing because in many other textual collections (e.g., Internet homepages, company databases) human-generated indexes may be unavailable or incomplete. Thus, the performance of the object filtering process could vary widely.

[1] 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.

In addition, some critiques claimed that because the object filters were created manually in the first place, the so-called system-generated thesauri are not entirely ''automatic.''

- *Automatic indexing only*

In order to develop a truly system-generated (automatic) thesaurus, we proceeded to a second experiment which included only automatic indexing. No object filters were used in this experiment. Multiple-word phrases were formed using a revised automatic indexing procedure. The rationale behind the revision was mainly based on suggestions made by various domain experts in our previous research (e.g., Russian computing researchers [Chen & Lynch, 1992], biologists [Chen et al., 1995]). In addition to a stop-word list of 536 functional words (including most of the prepositions), we also generated a list of 2,502 stop-verbs (i.e., words that serve only as verbs, e.g., generate, write, etc.). After removing instances of stop words and stop verbs from the abstracts, we found that a significant portion of the remaining words were adjectives and nouns. By performing term-phrase formation using two and three adjacent words, we were able to obtain a significant number of noun phrases, with one or two adjectives and a noun, a popular form for subject indexes. Because single-word terms often lacked specific meanings (e.g., system, model, tool) and included a significant amount of noise, we removed all single-word terms for further analysis. Thus, the thesaurus generated by adopting the revised automatic indexing process contained only two-word and three-word descriptors. (The thesaurus generated by using both object filtering and automatic indexing contained one-word terms and sometimes even four-word and five-word terms).

B. Co-Occurrence Analysis

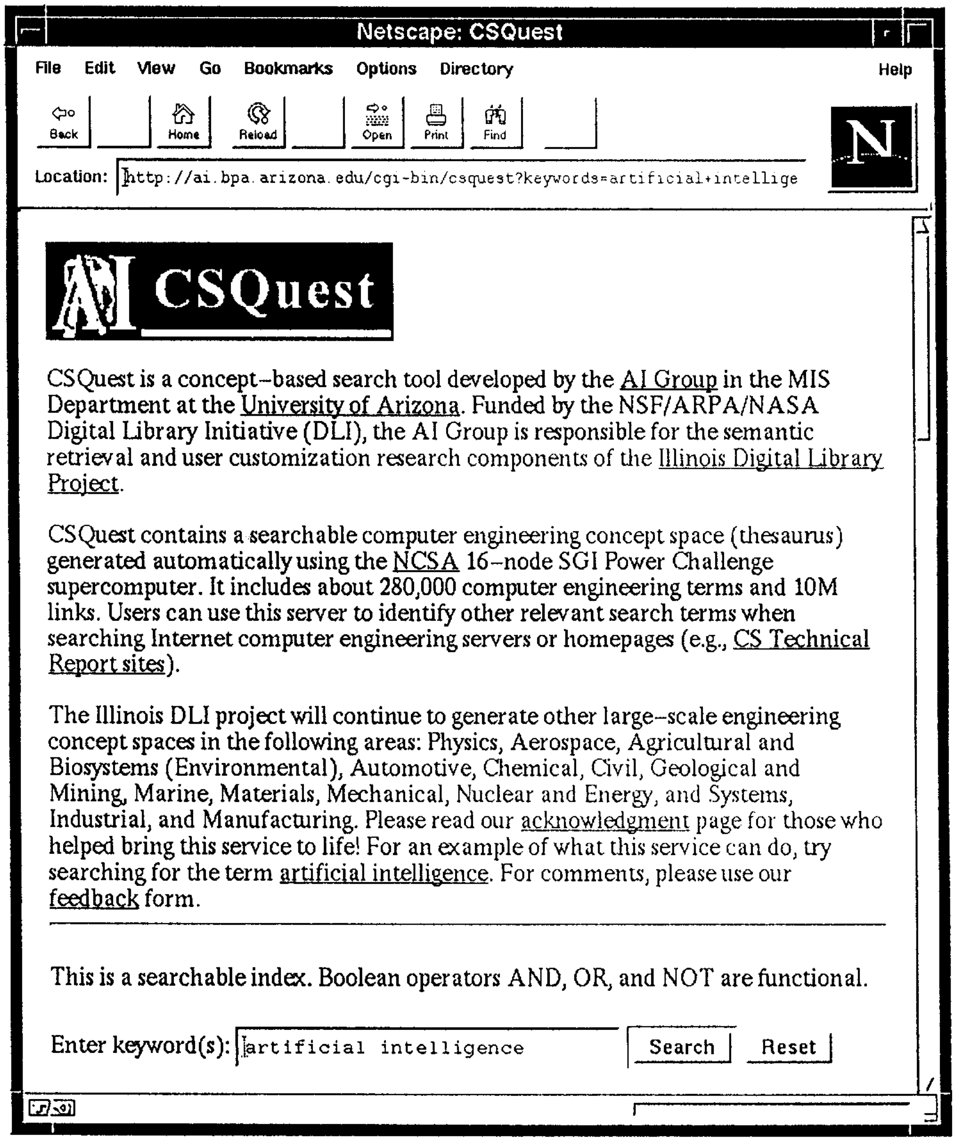
After terms were identified for each document, we proceeded to the co-occurrence analysis phase. Terms were weighted based on a revised ''term frequency'' and ''inverse document frequency'' measure. We then performed term co-occurrence analysis based on an asymmetric ''cluster similarity function'' developed by Chen and Lynch (1992). For algorithmic details about the co-occurrence analysis algorithm adopted in this project, readers are referred to Chen et al. (1995). (The co-occurrence analysis algorithm was unchanged in this experiment.)

The two sets of indexing results generated (i.e., a combined object filtering and automatic indexing method versus an automatic indexing only method) were analyzed using the above co-occurrence algorithm. After applying an empirically-determined co-occurrence threshold of 0.005, we were able to remove a significant portion of the less relevant co-occurrence pairs. The system-generated thesaurus (co-occurrence list), by using the combined method, contained 206,738 terms and 4,498,665 weighted, co-occurrence relationships (links). Our analysis revealed that about 98% of these terms were from object filters to which the various thresholds had been applied. Most of the automatic indexing terms did not survive the thresholds. Using automatic indexing alone, the system-generated thesaurus contained 683,762 terms and 61,328,941 relationships. We believe the number of terms for the second system-generated thesaurus was larger because of the term-phrase formation process. The human-generated INSPEC subject thesaurus contains about 15,700 terms: 7,700 are the preferred terms and 8,000 are the lead-in terms.

C. Parallel Computing

Due to the size of our collection (400,000+ abstracts, 2 GBs), generating the co-occurrence list on serial machines became problematic. In an earlier experiment involving 7,000+ abstracts, the entire object filtering, automatic indexing, and co-occurrence analysis process took about 50 minutes on a DEC Alpha 2100/600 workstation (200 MHz, 128-MB RAM). Our estimation of performing the same procedure for the 400,000+ collection on the same machine was about 30 days (CPU cycles). In order to alleviate this computational problem, an experiment was conducted recently on the NCSA (National Center for Supercomputing Applications) SGI Power Challenge shared-memory multiprocessor supercomputer (16 MIPS R8000 processors, with a total shared memory of 4 GBs). Using a *data parallel* strategy (i.e., allocating smaller, independent data files to different processors for processing), we were able to use all 16 processors in parallel, resulting in a significant speed-up. The computing time for producing two system-generated computing engineering thesauri was about 24 hours using the combined methods compared with 36 hours using automatic indexing only. Ninety percent of the computation was performed for co-occurrence analysis. (If fact, this was the largest single user of the NCSA Power Challenge supercomputer during the 2-week period.)

We believe, as large-scale digital libraries continue to proliferate and the tasks of searching a large information space become more overwhelming due to the search/indexing uncertainty, automatic thesaurus generation (a form of knowledge/concept discovery) will become essential to many textual applications. And as supercomputing evolves into newer and more diverse National and Grand Challenge applications, we see a clear trend of performing large-scale digital library analyses using the increasingly friendly and affordable supercomputing resources. This view was echoed in the recent ''Report on Workshop on High Performance Computing and Communications for Grand Challenge Applications: Computer Vision, Speech and Natural Language Processing, and Artificial Intelligence'' (Wah, 1993) and is popular among many information science researchers (Couvreur, Benzel, Miller, & Zeitler, 1994; Rasmussen, 1991).



Location: http://ai.bpa.arizona.edu/cgi-bin/csquest?keywords=artificial+intellige

AI CSQuest

CSQuest is a concept-based search tool developed by the AI Group in the MIS Department at the University of Arizona. Funded by the NSF/ARPA/NASA Digital Library Initiative (DLI), the AI Group is responsible for the semantic retrieval and user customization research components of the Illinois Digital Library Project.

CSQuest contains a searchable computer engineering concept space (thesaurus) generated automatically using the NCSA 16-node SGI Power Challenge supercomputer. It includes about 280,000 computer engineering terms and 10M links. Users can use this server to identify other relevant search terms when searching Internet computer engineering servers or homepages (e.g., CS Technical Report sites).

The Illinois DLI project will continue to generate other large-scale engineering concept spaces in the following areas: Physics, Aerospace, Agricultural and Biosystems (Environmental), Automotive, Chemical, Civil, Geological and Mining, Marine, Materials, Mechanical, Nuclear and Energy, and Systems, Industrial, and Manufacturing. Please read our acknowledgment page for those who helped bring this service to life! For an example of what this service can do, try searching for the term artificial intelligence. For comments, please use our feedback form.

This is a searchable index. Boolean operators AND, OR, and NOT are functional.

Enter keyword(s): artificial intelligence

FIG. 1. A user entered ''artificial intelligence'' in the CSQuest thesaurus search window. All terms matched with ''artificial intelligence'' were displayed in ranked order based on WAIS indexing.

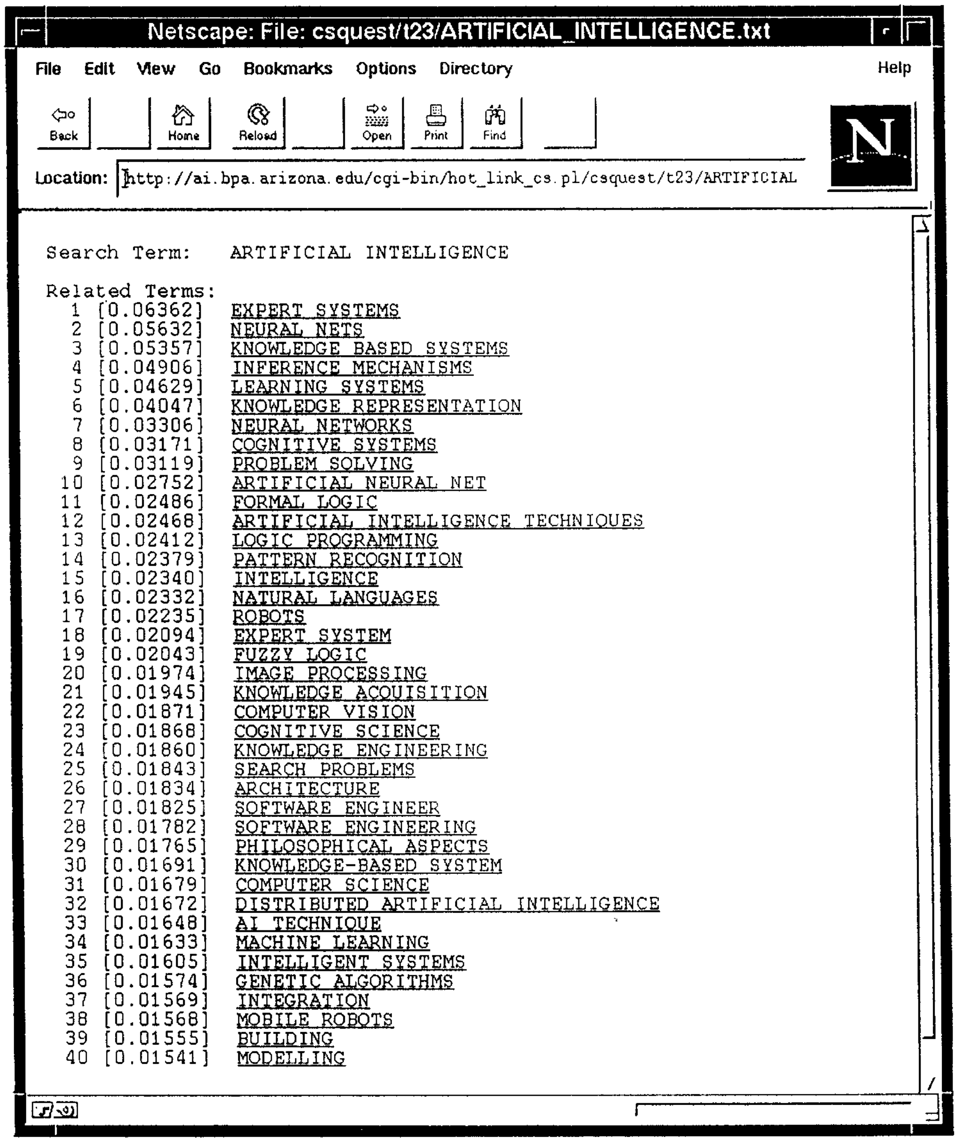
The system-generated thesaurus using the combined object filtering and automatic indexing method was recently incorporated into a WWW server. The server, which is called *CSQuest,* allows searchers to significantly increase the ''variety'' in their search terms by picking system-generated terms. For system performance reasons, we list only the top 40 ranked terms for any given search term. The server can be accessed at: http://ai.bpa.arizona.edu/html/csquest/. Selected screen dumps of the CSQuest server are shown in Figures 1 and 2.

4. A Concept Association Experiment

In order to examine the ability of the system-generated thesauri in generating meaningful (relevant) ''variety,'' we conducted a user evaluation experiment, using the human-generated INSPEC thesaurus as a benchmark for comparison.

4.1. Experimental Design

A two-phase experiment involving nine faculty and graduate student subjects affiliated with an Information Systems Department was conducted. All subjects were familiar with the subject areas of Artificial Intelligence and Databases. Based on a list of candidate terms generated by two faculty members, we selected 12 test descriptors (six terms in Artificial Intelligence and six terms in Databases) that occurred in both system-generated thesauri as well as in the human-generated INSPEC subject


Netscape: File: csquest/t23/ARTIFICIAL_INTELLIGENCE.txt
File Edit View Go Bookmarks Options Directory Help
Back
Home
Reload
Open
Print
Find
N

Location: http://ai.bpa.arizona.edu/cgi-bin/hot_link_cs.pl/csquest/t23/ARTIFICIAL

```
Search Term:   ARTIFICIAL INTELLIGENCE

Related Terms:
  1 [0.06362]  EXPERT SYSTEMS
  2 [0.05632]  NEURAL NETS
  3 [0.05357]  KNOWLEDGE BASED SYSTEMS
  4 [0.04906]  INFERENCE MECHANISMS
  5 [0.04629]  LEARNING SYSTEMS
  6 [0.04047]  KNOWLEDGE REPRESENTATION
  7 [0.03306]  NEURAL NETWORKS
  8 [0.03171]  COGNITIVE SYSTEMS
  9 [0.03119]  PROBLEM SOLVING
 10 [0.02752]  ARTIFICIAL NEURAL NET
 11 [0.02486]  FORMAL LOGIC
 12 [0.02468]  ARTIFICIAL INTELLIGENCE TECHNIQUES
 13 [0.02412]  LOGIC PROGRAMMING
 14 [0.02379]  PATTERN RECOGNITION
 15 [0.02340]  INTELLIGENCE
 16 [0.02332]  NATURAL LANGUAGES
 17 [0.02235]  ROBOTS
 18 [0.02094]  EXPERT SYSTEM
 19 [0.02043]  FUZZY LOGIC
 20 [0.01974]  IMAGE PROCESSING
 21 [0.01945]  KNOWLEDGE ACQUISITION
 22 [0.01871]  COMPUTER VISION
 23 [0.01868]  COGNITIVE SCIENCE
 24 [0.01860]  KNOWLEDGE ENGINEERING
 25 [0.01843]  SEARCH PROBLEMS
 26 [0.01834]  ARCHITECTURE
 27 [0.01825]  SOFTWARE ENGINEER
 28 [0.01782]  SOFTWARE ENGINEERING
 29 [0.01765]  PHILOSOPHICAL ASPECTS
 30 [0.01691]  KNOWLEDGE-BASED SYSTEM
 31 [0.01679]  COMPUTER SCIENCE
 32 [0.01672]  DISTRIBUTED ARTIFICIAL INTELLIGENCE
 33 [0.01648]  AI TECHNIQUE
 34 [0.01633]  MACHINE LEARNING
 35 [0.01605]  INTELLIGENT SYSTEMS
 36 [0.01574]  GENETIC ALGORITHMS
 37 [0.01569]  INTEGRATION
 38 [0.01568]  MOBILE ROBOTS
 39 [0.01555]  BUILDING
 40 [0.01541]  MODELLING
```

FIG. 2. CSQuest displayed 40 ranked system-generated thesaurus terms related to ‘‘artificial intelligence.’’ Each term can be clicked on for more related terms.

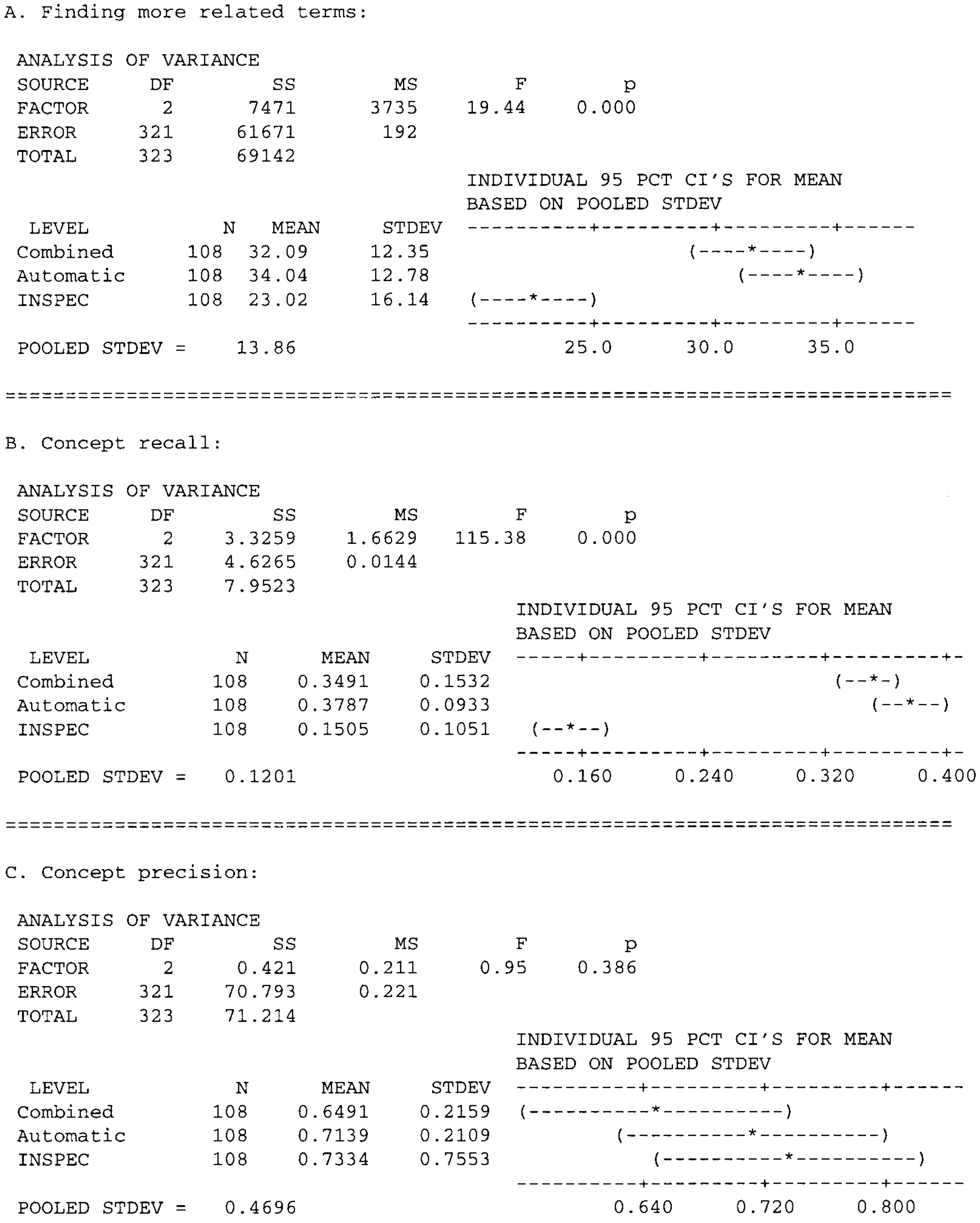
thesaurus. We adopted an experimental design similar to those used in human memory association experiments (Anderson, 1985) and in thesaurus evaluation studies (Chen et al., 1995). A *recall phase,* followed by a *recognition phase,* was performed in order.

In Phase 1 (*Recall Phase*), each subject was asked to generate through a free association process as many related terms as possible in response to each test descriptor presented. This phase of the experiment called upon subjects' memory recall. In Phase 2 (*Recognition Phase*), experimenters created lists of associated terms in random order for subjects to evaluate with regard to their relevance to the test descriptor. Included in each list were the 30 highest-ranked terms suggested by each of the system-generated thesauri and all terms suggested by the INSPEC thesaurus. (We chose only the top 30 terms because of operational issues in user evaluation.) The nine subjects were then asked to evaluate each thesaurus-suggested term according to a Likert-like scale: ‘‘Irrelevant,’’ ‘‘Somewhat Relevant,’’ ‘‘Very Relevant.’’ Terms considered too general were to be ranked as ‘‘Irrelevant.’’ This phase of the experiment called upon the subjects' ability to recognize relevant terms. The complete experiment lasted between 1.25 hours and 2.5 hours for each subject.

4.2. Experimental Results

- *Finding more related terms*

By counting the number of terms generated by the subjects themselves in the *Recall Phase* of the experi-



```
A. Finding more related terms:

 ANALYSIS OF VARIANCE
 SOURCE     DF        SS        MS         F        p
 FACTOR      2      7471      3735     19.44    0.000
 ERROR     321     61671       192
 TOTAL     323     69142
                                      INDIVIDUAL 95 PCT CI'S FOR MEAN
                                      BASED ON POOLED STDEV
  LEVEL           N   MEAN     STDEV  ----------+---------+---------+------
 Combined      108  32.09     12.35                     (----*----)
 Automatic     108  34.04     12.78                         (----*----)
 INSPEC        108  23.02     16.14   (----*----)
                                      ----------+---------+---------+------
 POOLED STDEV =    13.86                      25.0      30.0      35.0

==============================================================================

B. Concept recall:

 ANALYSIS OF VARIANCE
 SOURCE     DF        SS        MS         F        p
 FACTOR      2    3.3259    1.6629   115.38    0.000
 ERROR     321    4.6265    0.0144
 TOTAL     323    7.9523
                                          INDIVIDUAL 95 PCT CI'S FOR MEAN
                                          BASED ON POOLED STDEV
  LEVEL            N      MEAN     STDEV  -----+---------+---------+---------+-
 Combined        108    0.3491    0.1532                            (--*-)
 Automatic       108    0.3787    0.0933                               (--*--)
 INSPEC          108    0.1505    0.1051   (--*--)
                                          -----+---------+---------+---------+-
 POOLED STDEV =   0.1201                     0.160     0.240     0.320     0.400

==============================================================================

C. Concept precision:

 ANALYSIS OF VARIANCE
 SOURCE     DF        SS        MS         F        p
 FACTOR      2     0.421     0.211     0.95    0.386
 ERROR     321    70.793     0.221
 TOTAL     323    71.214
                                          INDIVIDUAL 95 PCT CI'S FOR MEAN
                                          BASED ON POOLED STDEV
  LEVEL            N      MEAN     STDEV  ----------+---------+---------+------
 Combined        108    0.6491    0.2159  (----------*----------)
 Automatic       108    0.7139    0.2109          (----------*----------)
 INSPEC          108    0.7334    0.7553             (----------*----------)
                                          ----------+---------+---------+------
 POOLED STDEV =   0.4696                          0.640     0.720     0.800
```

FIG. 3. ANOVA analysis results for finding more terms, concept recall, and concept precision.

ment, together with the system-suggested terms marked by the subjects as either somewhat relevant or very relevant, we were able to tabulate and analyze the contribution of each thesaurus to the search terms. An analysis of variance (ANOVA) procedure was conducted for the number of terms generated, using the MINITAB statistical analysis package (Ryan, Joiner, & Ryan, 1985), followed by a two-sample t-test to determine the differences in means. The results are summarized in Figure 3. On average, the two system-generated thesauri were able to suggest significantly more terms (the combined method: 32.09 terms and the automatic indexing method: 34.04 terms) than the INSPEC Thesaurus (23.02 terms) (p = 0.000). The two-sample t-test revealed statistical differences (at a level of significance of 10%) between each of the system-generated thesauri and the INSPEC Thesaurus (combined vs. INSPEC: $p = 0.000$; automatic vs. INSPEC: $p = 0.000$), but no statistical difference between the two system-generated thesauri themselves, $p = 0.257$).

- *Concept recall and concept precision*

For analysis of results from the concept association experiment, we utilized *concept recall* and *concept precision* for evaluation, rather than the *document* recall and precision measures typically used in information science research. Rather than examining the number of relevant documents retrieved, we counted the number of relevant terms (concepts) generated by the thesaurus. These measures, which are grounded mostly in cognitive psychology research, have also been adopted in thesaurus evaluation studies (Chen & Lynch, 1992; Chen et al., 1995). They were computed as follows:

$$\textit{Concept Recall} = \frac{\text{Number of Retrieved Relevant Concepts}}{\text{Number of Total Relevant Concepts}}$$

$$\textit{Concept Precision} = \frac{\text{Number of Retrieved Relevant Concepts}}{\text{Number of Total Retrieved Concepts}}$$

Total Relevant Concepts represented the target set of concepts that could be obtained through user-thesaurus interaction and included all concepts generated by the subjects in Phase 1, as well as those additional unique concepts selected as relevant by the subjects from the two system-generated thesauri and the INSPEC subject thesaurus in Phase 2. *Total Retrieved Concepts* represented the number of relevant concepts suggested by each thesaurus (30 for each system-generated thesaurus and a number for the INSPEC thesaurus that varied). *Retrieved Relevant Concepts* represented the number of concepts for each thesaurus judged ‘‘Very Relevant’’ or ‘‘Somewhat Relevant’’ by the subjects. ANOVA tests and two-sample t-tests were performed for *concept recall* and *concept precision.*

As shown in Figure 3, concept recall for the system-generated thesaurus using automatic indexing only (37.9%) was significantly better than that for either the INSPEC subject thesaurus (15.1%) or the system-generated thesaurus using the combined method (34.9%) (the overall difference was significant at $p = 0.000$). Two-sample t-tests confirmed the significance of the difference between the recall value (at the level of significance of 10%) for combined vs. automatic ($p = 0.088$), combined vs. INSPEC ($p = 0.000$), and automatic vs. INSPEC ($p = 0.000$).

The greater recall values for the system-generated thesauri can be attributed to the systems’ ability to identify the contextual associations between concept pairs in a large collection of domain-specific documents. The related concepts suggested by the system-generated thesauri were often more comprehensive and up-to-date. The difference between recall values for the two system-generated thesauri ($p = 0.088$) was somewhat surprising. We believe this is because some terms identified by object filtering were considered too general by many subjects and thus rated as irrelevant. This problem also occurred for the INSPEC subject thesaurus.

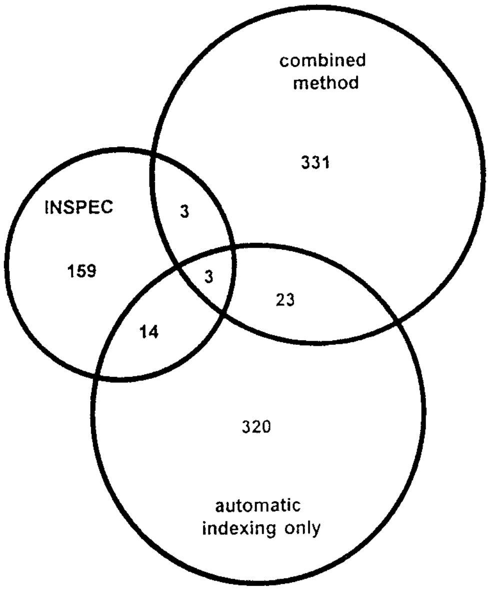


FIG. 4. Intersection of concept associations from three sources.

Concept precision for each of the system-generated thesauri was slightly lower than that for the INSPEC subject thesaurus (64.9 and 71.4% vs. 73.3%). However, this result was not statistically significant overall ($p = 0.386$). We are encouraged that the system-generated thesauri were able to obtain such a high level of precision, comparable to that of the human-generated thesaurus. That the precision for INSPEC was not 100% can be explained by the fact that although terms in a manually generated thesaurus are carefully selected to represent a limited number of highly relevant terms, subjects typically deemed broader or parent terms as irrelevant (i.e., not appropriate for search), which lessened the number of potentially relevant terms included in the set suggested.

4.3. Discussion

A further analysis was performed following the experiment to examine the similarity (intersection) among the terms suggested by the three sources. Our goal was to explore the prospect of integrating multiple thesauri (human-generated and system-generated) for generating ‘‘variety’’ in search.

The total number of terms suggested by all three sources in response to the 12 terms specified was 899. Each system-generated thesaurus suggested 360 terms (30 terms each for the 12 concepts) and the INSPEC subject thesaurus suggested 179 terms. As shown in Figure 4, of the 899 associations suggested by the three sources, only three of the terms were suggested by all three sources. Seventeen terms were suggested by both the INSPEC thesaurus and the automatic-indexing-generated thesaurus, and 26 terms were suggested by both system-generated thesauri. We found only a small overlap of six terms between the INSPEC thesaurus and the combined-method-generated thesaurus. Illustrating an exam-

combined method	automatic indexing only	INSPEC thesaurus
#object-oriented DBMS	object-oriented database system	#database management systems
database tools	data base system	computer applications
query learning	object-oriented database management	file-organization
query processing algorithm	#database management system	object-oriented methods
object-oriented programming approach	management system	object-oriented programming
database system design	programming language	
deductive DBMS	object-oriented database design	
database management technology	complex object	
data integrity problems	database model	
data structures course	database language	
knowledge based systems development	database schema	
object-oriented database systems	object-oriented database programming	
software toolset	object-oriented paradigm	
#query languages	database programming language	
distributed DBMS	object identity	
object-oriented data structure	object-oriented database model	
concurrency control algorithm	query optimization	
object-oriented database technology	database technology	
object-oriented layer	object-oriented database language	
international connections	integrity constraint	
data modem	object-oriented programming language	
database schema design	object oriented database	
programming equipment	object-oriented system	
relational data base	deductive object-oriented database	
object oriented database management system	international conference	
formal specification language	schema evolution	
complex object databases	oriented database system	
database theory	graphical user interface	
software reusable	#object-oriented DBMS	
data modeling concepts	#query languages	

Common terms are marked by #.

FIG. 5. Terms suggested for the query term ''object-oriented databases'' by the three thesauri.

ple of this pattern of overlap, Figure 5 displays the terms suggested by each source for the query term ''object-oriented databases.'' For this query term, only three terms were common among the three sources. (In our computation, plural and singular forms of a term were counted as the same, e.g., database management system(s).)

It is evident, from the example in Figure 5, that despite this seemingly small overlap of common terms, all three sources actually produced concepts which have a significant ''semantic'' overlap (but differ syntactically). The system-generated thesaurus using the combined method produced object-oriented programming approach, object-oriented database systems, object-oriented data structures, object-oriented layer, object-oriented database technology, etc. The system-generated thesaurus using automatic indexing only produced object-oriented database management, object-oriented database design, object-oriented database programming, object-oriented paradigm, object-oriented database model, etc.; the INSPEC thesaurus generated object-oriented methods and object-oriented programming. Semantic overlap is particular apparent for the two system-generated thesauri.

In summary, we believe the syntactic and semantic variations provided by the three vocabulary sources could potentially provide a useful and system-aided way of generating ''variety'' for search terms, especially by allowing search terms that match well with the system-generated thesauri to ''dock'' onto a more selective list of indexing terms that match well with the INSPEC subject thesaurus. (''Docking'' is a concept for alleviating search uncertainty that is regarded highly by information science researchers [Bates, 1986]).

5. Conclusion and Discussion

Using the INSPEC thesaurus as the benchmark for comparison, system-generated computer engineering thesauri have demonstrated their potential usefulness for suggesting search terms. We are convinced that multiple interfaces and multiple vocabulary search aids are necessary for effective concept-based search across multiple large-scale repositories and domains.

Our current work in the ongoing Illinois digital library initiative project mainly involves: (1) Creating system-generated thesauri for other major engineering domains (roughly in the following order: Chemical, materials, systems and industrial manufacturing, mechanical, aerospace, automatic, civil, agricultural and biosystems, geological and mining, marine, and nuclear and energy) using the 48-processor SGI Power Challenge Array and 64-processor Convex Exemplar super-computers at NCSA (a 5,000,000+ collection of abstracts in all engineering domains, 1986–1995, has been provided by the Compendex database); and (2) developing robust graph matching and traversal algorithms for cross-domain term switching (Chen & Ng, 1995).

6. Acknowledgments

This project was supported mainly by the following grants:

- NSF/ARPA/NASA Digital Library Initiative, IRI-9411318, 1994–1998 (B. Schatz, H. Chen, et al., ‘‘Building the Interspace: Digital Library Infrastructure for a University Engineering Community’’);
- NSF, CISE, IRIS, IRI-9525790, 1995–1998 (H. Chen, ‘‘Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach’’);
- NSF CISE Research Initiation Award, IRI-9211418, 1992–1994 (H. Chen, ‘‘Building a Concept Space for an Electronic Community System’’);
- NSF CISE Special Initiative on Coordination Theory and Collaboration Technology, IRI-9015407, 1990–1993 (B. Schatz, ‘‘Building a National Collaboratory Testbed’’);
- AT&T Foundation Special Purpose Grants in Science and Engineering, 1994–1995 (H. Chen); and
- National Center for Supercomputing Applications (NCSA), High-performance Computing Resources Grants, IRI-9500001N, 1994–1996 (H. Chen).

We would also like to thank Dr. Anindya Datta, Pauline Cochrane, and Dr. Ann Bishop for their valuable suggestions; Jim Ashling of the Institution of Electrical Engineers (IEE, vendor of the INSPEC database) for providing the INSPEC thesaurus for our experiment; and Dr. Larry Smarr, Dr. Melanie Loots, and Mike Welge at NCSA for their kind assistance.

References

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: W. H. Freeman and Company.

Bates, M. J. (1977). System meets user: Problems in matching subject search terms. *Information Processing and Management, 13*(6), 367–368.

Bates, M. J. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science, 37,* 357–376.

Bellamy, L. M., & Bickham, L. (1989, Winter). Thesaurus development for subject cataloging. *Special Libraries, 80,* 9–15.

Chamis, A. Y. (1991). *Vocabulary control and search strategies in online searching.* New York: Greenwood Press.

Chan, L. M. (1986). *Library of Congress subject headings: Principles and application* (2nd ed.). Littleton, CO: Libraries Unlimited.

Chaplan, M. A. (1995). Mapping Laborline thesaurus terms to Library of Congress subject headings: Implications for vocabulary switching. *Library Quarterly, 65*(1), 39–61.

Chen, H. (1994). Collaborative systems: Solving the vocabulary problem. *IEEE Computer,* [Special issue on computer-supported cooperative work (CSCW)], *27*(5), 58–66.

Chen, H., Hsu, P., Orwig, R., Hoopes, L., & Nunamaker, J. F. (1994). Automatic concept classification of text from electronic meetings. *Communications of the ACM, 37*(10), 56–73.

Chen, H., & Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics, 22*(5), 885–902.

Chen, H., Lynch, K. J., Basu, K., & Ng, D. T. (1993). Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert, Special Series on Artificial Intelligence in Text-based Information Systems, 8*(2), 25–34.

Chen, H., & Ng, D. T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science, 46,* 348–369.

Chen, H., Schatz, B. R., Yim, T., & Fye, D. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46,* 175–193.

Couvreur, T. R., Benzel, R. N., Miller, S. F., & Zeitler, D. N. (1994). An analysis of performance and cost factors in searching large text databases using parallel search systems. *Journal of the American Society for Information Science, 45,* 443–464.

Crouch, C. J. (1990). An approach to the automatic construction of global thesauri. *Information Processing and Management, 26*(5), 629–640.

Crouch, C. J., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval,* Copenhagen, Denmark, June 21–24, 1992 (pp. 77–88).

Everitt, B. (1980). *Cluster analysis* (2nd ed.). London: Heinemann Educational Books.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964–971.

Gomez, L. M., & Lochbaum, C. C. (1984). People can retrieve more objects with enriched key-word vocabularies. But is there a human performance cost? In B. Shackel (Ed.), *Human–computer interaction—Interact ’84.* (pp. 257–261). Amsterdam: North-Holland.

Gomez, L. M., Lochbaum, C. C., & Landauer, T. K. (1990). All the right words: Finding what you want as a function of the richness of indexing vocabulary. *Journal of the American Society for Information Science, 41,* 547–559.

Good, M. D., Whiteside, J. A., Wixon, D. R., & Jones, S. J. (1984). Building a user-derived interface. *Communications of the ACM, 27*(10), 1032–1043.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. (1983). *Building expert systems.* Reading, MA: Addison-Wesley.

Jacoby, J., & Slamecka, V. (1962). *Indexer consistency under minimal conditions.* Bethesda, MD: Documentation, Inc.

Kim, Y. W., & Kim, J. H. (1990). A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation, 46,* 113–116.

Knapp, S. D. (1984). Creating BRS/TERM, a vocabulary database for searchers. *Database, 7*(4), 70–75.

Lindberg, D. A., & Humphreys, B. L. (1990). The UMLS knowledge sources: Tools for building better user interface. In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care,* November, 4–7 1990 (pp. 121–125). Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

McCray, A. T., & Hole, W. T. (1990). The scope and structure of the first version of the UMLS semantic network. In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care,* November, 4–7 1990, (pp. 126–130). Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Niehoff, R. T. (1976). Development of an integrated energy vocabulary and the possibilities for on-line subject switching. *Journal of the American Society for Information Science, 27*(1), 3–17.

Niehoff, R. T., & Kwansy, S. (1979). The role of automated subject switching in a distributed information network. *Online Review, 3*(2), 181–194.

Petersen, T. (1983). The AAT: A model for the restructuring of LCSH. *Journal of Academic Librarianship, 9*(4), 207–210.

Petersen, T. (1990). Developing a new thesaurus for art and architecture. *Library Trends, 38*(4), 644–658.

Piternick, A. B. (1984). Searching vocabularies: A developing category of online search tools. *Online Review, 8*(5), 441–449.

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms.* Englewood Cliffs, NJ: Prentice Hall.

Rasmussen, E. M. (1991). Introduction: Parallel processing and information retrieval. *Information Processing and Management, 27*(4), 255–263.

Ryan, B. F., Joiner, B. L., & Ryan, T. A. (1985). *MINITAB Handbook* (2nd ed.). Boston, MA: PWS-KENT Publishing Company.

Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison-Wesley Publishing Company, Inc.

Stevens, M. E. (1965). *Automatic indexing: A state-of-the-art report.* Washington, DC: U.S. Government Printing Office.

Wah, B. (1993). Report on workshop on high performance computing and communications for grand challenge applications: Computer vision, speech and natural language processing, and artificial intelligence. *IEEE Transactions on Knowledge and Data Engineering, 5*(1), 138–154.