Automatic Thesaurus Generation for an Electronic Community System

Hsinchun Chen, Tak Yim, and David Fye
University of Arizona, Management Information Systems Department, Karl Eller Graduate School of Management, McClelland Hall 430Z, Tucson, AZ 85721. E-mail for Hsinchun Chen: hchen@bpa.arizona.edu

Bruce Schatz
University of Illinois, Graduate School of Library and Information Science, National Center for Supercomputing Applications, Beckman Institute, 405 N. Mathews Avenue, Urbana, IL 61801. E-mail: bschatz@ncsa.uiuc.edu

This research reports an algorithmic approach to the automatic generation of thesauri for electronic community systems. The techniques used included *term filtering, automatic indexing,* and *cluster analysis.* The testbed for our research was the *Worm Community System,* which contains a comprehensive library of specialized community data and literature, currently in use by molecular biologists who study the nematode worm *C. elegans.* The resulting worm thesaurus included 2709 researchers' names, 798 gene names, 20 experimental methods, and 4302 subject descriptors. On average, each term had about 90 weighted neighboring terms indicating *relevant* concepts. The thesaurus was developed as an online search aide. We tested the worm thesaurus in an experiment with six worm researchers of varying degrees of expertise and background. The experiment showed that the thesaurus was an excellent "memory-jogging" device and that it supported learning and serendipitous browsing. Despite some occurrences of obvious noise, the system was useful in suggesting relevant concepts for the researchers' queries and it helped improve concept *recall.* With a simple browsing interface, an automatic thesaurus can become a useful tool for online search and can assist researchers in exploring and traversing a dynamic and complex electronic community system.

Introduction

Electronic community systems, which encode a research community's information and knowledge and provide an online environment to support the manipulation of that knowledge, have drawn significant attention recently due to the rapid proliferation and advancement of computing, databases, and telecommunication technologies. An electronic community system helps researchers in the community function more efficiently and effectively by allowing them to browse the available knowledge easily, record their own knowledge for others to use, and form interrelationships between concepts (Schatz, 1991/1992).

The *Worm Community System* (WCS) (Pool, 1993; Schatz, 1991/1992, 1993), a recent implementation of electronic community systems technology, has been developed as part of the NSF *collaboratory* effort for the community of molecular biologists who study the nematode worm *Caenorhabditis elegans.* Molecular biology is a largely data-driven experimental science and, due to such efforts as the Human Genome Initiative, its data accumulation is growing rapidly and being stored in databases. Communities in molecular biology often form around organisms, rather than techniques or problems. The "worm" has become a primary model organism for genome research and will likely become the first to be completely sequenced.

Despite the usefulness of database technologies for community systems, researchers are often overwhelmed by the amount of current information, the subject and computer knowledge required to access this information, and the constant influx of new information. This results in what is termed *information overload.* A second difficulty associated with information retrieval and information sharing is the classical *vocabulary* problem, which results from the diversity of expertise and backgrounds of system users. For an operational electronic community system like the WCS, potential users vary from expert worm biologists to novices, and from senior worm insiders to community outsiders (e.g., fly biologists). These users often do not share the same vocabularies and they may experience difficulties using system-specific query terms.

Received June 21, 1993; revised March 14, 1994; accepted May 2, 1994.

The *fluidity* of concepts and vocabularies in the scientific domains (especially in genome research) (Frenkel, 1991) further complicates the retrieval issue. A scientific concept may be perceived differently by different researchers and it may also convey different meanings at different times. To address the *information overload* and the *vocabulary* problems in information retrieval, our research presents an algorithmic approach to the automatic generation of a domain-specific thesaurus.

Because the research in worm biology is so new and specialized, there exists no thesaurus in the field and the option of compiling an expert-generated thesaurus was infeasible due both to the time and effort required and to the difficulty of updating such a thesaurus in a timely manner. Our proposed approach generated a domain-specific thesaurus automatically by analyzing stored documents using externally acquired controlled term (keyword) lists, automatic indexing techniques, and statistics-based cluster analysis algorithms (Chen & Lynch, 1993). The resulting thesaurus captured domain-specific concepts and their weighted, relevant relationships, and it allowed automatic, periodic update of its vocabularies and relationships. With a simple browsing interface, users can access the WCS using their own vocabularies and consult the semantics-rich thesaurus for other similar concepts. The high-level *concept space* represented in the thesaurus also will allow users to search the system's large amount of knowledge and literature in a more manageable and timely manner (i.e., it addresses the information overload problem).

The Worm Community System and Information Retrieval Problems

The Human Genome Initiative and the Worm Community System

In recent years the Human Genome Initiative has sparked significant interest in the biology, biomedicine, and genetics research communities and it has also called for advanced computing technologies to support such research activities (Courteau, 1991; Frenkel, 1991; Lander, Langridge, & Saccocio, 1991). According to Courteau, the Human Genome Project "will generate more data than any single project to date in biology." It will list the map and location of the entire human genome and the genomes of other model organisms along with the associated literature and knowledge accumulated during the project's scientific discovery process. Because of the need to store, manage, analyze, and disseminate such a diverse and vast amount of information and knowledge, computing technologies that can support efficient and effective storage and retrieval of information, foster seamless distributed scientific collaboration, and facilitate timely information dissemination and sharing are needed.

In addition to various genome databases created for various domains and model organisms (e.g., coli, yeast, fly, worm, mouse, and human) (Courteau, 1991), electronic community systems (ECS) which enable researchers of a scientific community to enter and share community knowledge and findings in a timely manner and in a distributed environment have been proposed and implemented. An advantage of this type of system over traditional databases is that an ECS enables users to add new data from their research, annotate entries from others' research, and indicate authorizations for users to either view or annotate their own data (Courteau, 1991). Another novel characteristic of the ECS is its ability to handle a wide variety of community knowledge. It encodes formal and informal knowledge and their interrelationships within a subject domain and allows easy manipulation of such knowledge (Schatz, 1993). To "live effectively within a community, one must have available both the formal archival material and the informal transient folklore" (Schatz, 1993). An ECS is much like an electronic library where users can browse for relevant information, filter out the information they do not currently need, and share the data that they have found.

The Worm Community System (WCS), which is a major NSF-funded collaboratory project, has been considered a model electronic community system. This experiment in building an electronic community system for the *C. elegans* researchers offers traditional database functionalities along with literature, informal information and research lore, mapping programs, and graphics, and the ability for users to browse, share, and filter a large amount of timely worm community knowledge. The system is intended to serve the entire community of worm biologists and other related biology and biomedical community members (Courteau, 1991; Schatz, 1993).

The current WCS runs under X-Windows on Unix machines and can also be used remotely from X-Windows on Macs, PCs, SUNs, DECs, and X-Windows terminals. Currently, the data existing in the WCS include: (1) archival data such as gene descriptions, a physical map, DNA sequences, a cell list, and cell lineage; (2) formal and informal literature such as the worm (reference) book, journal abstracts, *Worm Breeder's Gazette* (newsletter) articles, conference abstracts, lab directory, and community lore; and (3) analysis software (e.g., programs for displaying maps and sequences).

Information Retrieval (IR) Problems Encountered

As the WCS matures, many of the users (both experts and novices) have "mentioned fears of being overwhelmed by the amount of information in the annotations" (Star, 1991). Concerns included in the preliminary WCS field evaluation report (Star, 1991) were an overuse of annotations that would clutter up the system. Also, several users suggested that a filter should be used

on bulletin boards and annotations to eliminate less important information that tends to get into the database.

Chen and Lynch (1992) discussed the problem of information overload typically occurring in operational databases after years of use. For the WCS, there already exists a large amount of scientific knowledge of various forms and new scientific knowledge is being rapidly added on a regular basis. This acceleration of information gathering has resulted in difficulty in maintaining the database over time and made it hard for untrained users to perform online information retrieval (Frenkel, 1991). The information overload problem is complicated by the diversity of subject area knowledge, classification knowledge, and system knowledge exhibited by different users (Chen & Dhar, 1990, 1991). Simple browsing systems can potentially confuse and disorient a user, the *embedded digression problem,* and can cause the user to spend a great deal of time browsing while learning nothing specific, the *art museum phenomenon* (Carmel, Crawford, & Chen, 1992; Foss, 1989).

Further compounding the problem of information overload is the wide variety of knowledge contained in the scientific community, whether in formal documents such as literature abstracts and journal entries or in informal sources such as electronic mail and newsletters. The unstructured and varying formats of the scientific community knowledge create yet another cognitive demand for users of online databases.

The second problem, the *vocabulary problem,* has its roots in information science research. Previous research in indexing and subject access (Blair, 1986) and in human–computer interaction (Furnas, 1987) has shown that people tend to use different terms to describe a similar concept. Due to the unique backgrounds, training, and experiences of different users, the chance of two people using the same term to describe a concept is quite low and even the same person may use different terms to describe the same concept at different times (due to the learning process and the evolution of concepts).

Vocabulary is particularly problematic in a scientific community because of the diversity of specialized domains and the process of scientific discovery (especially in genome research). "Biology . . . involves concepts that are dynamic, or fluid, meaning that the phenomena under study, and the scientists' understanding of them, keep changing" (Frenkel, 1991). Experimental errors or approximations are common occurrences and definitions for concepts that follow these errors will evolve over time. According to Frenkel, the meanings of concepts "become better understood as more knowledge is accumulated and integrated." This novel characteristic of changing definitions over time must be implemented into the community system to make the system more flexible. Research that deals with an "old" concept must still be accessible by the users even though the terminology is no longer in common use.

For the WCS, we have identified three types of users: subject area experts; subject area novices; and outsiders (researchers from other disciplines). Keyword searching and the browsing interface of conventional databases are restricted in their ability to allow searchers to use their own vocabularies in search. For example, a molecular biologist researching bcl-2, the human gene which controls cell death, will be likely to experience great difficulty finding relevant information in the *Worm* database using bcl-2 (the gene similar to bcl-2 in *Worm* is ced-9). How to allow searchers who are not familiar with the specific subject area and terminology of a database to express queries using their own vocabularies is one of the most pressing research questions in information science.

An important direction of future WCS development efforts will involve a concept-based retrieval interface based on an automatically generated worm thesaurus and a system-aided thesaurus consultation process. The interface will allow searchers to access the worm database using their own vocabularies. This article reports findings regarding the first stage of this effort—automatic thesaurus generation.

An Algorithmic Approach to Automatic Thesaurus Generation

Our thesaurus project is mainly grounded in *knowledge discovery in databases,* an active research area which has drawn researchers from artificial intelligence, databases, and information science. Our proposed approach includes techniques from statistical analysis, automatic text processing, and knowledge-based system design. We discuss these techniques below in the context of earlier research.

Techniques: An Overview

"Knowledge discovery in databases" or knowledge "mining" is believed by many artificial intelligence and database researchers to be useful for resolving the information overload problem and for acquiring knowledge from databases. As Piatetsky-Shapiro (1989) remarked at the first workshop on knowledge discovery in the International Joint Conference on Artificial Intelligence, 1989:

> The growth in the amount of available databases far outstrips the growth of corresponding knowledge. This creates both a need and an opportunity for extracting knowledge from databases. Many recent results have been reported on extracting different kinds of knowledge from databases, including diagnostic rules, drug side effects, classes of stars, rules for expert systems, and rules for semantic query optimization. (*AI Week* March 15, 1990). At a recent NSF Invitational Workshop on the future of database research (Lagunita, CA, Feb. 1990), "knowledge mining" was among the top five research topics.

Sample areas in which knowledge discovery is applicable include: identifying patterns in frequent flyer databases and behaviors of credit-card holders, developing diagnostic expert systems from car trouble symptoms and problems found, and examining corporate intelligence reports and patterns of airline crashes or tax fraud (Piatetsky-Shapiro, 1989; Quint, 1991). Many major companies and government agencies are "learning" as much as possible from their databases. The knowledge discovered has often resulted in significant competitive advantage.

In this research, our aim was to apply an algorithmic approach to the generation of a robust knowledge base based on statistical correlation analysis of the concepts (knowledge) embedded in the documents of real-life, textual databases. The research output consisted of a thesaurus-like knowledge base, which can aid in concept-based information management and retrieval. This automatically generated thesaurus component, akin to a manually created thesaurus, can also play an important role in solving searchers' vocabulary problems during information retrieval.

In information science, use of a thesaurus or knowledge base for "intelligent" information retrieval has also drawn significant attention in recent years. There have been many attempts to capture experts' domain knowledge for information retrieval. For example, CoalSORT (Monarch & Carbonell, 1987), a knowledge-based interface, facilitates the use of bibliographic databases on coal technology. A semantic network, representing an expert's domain knowledge, embodies the system's intelligence. Fox's CODER system (Fox, 1987) consists of a thesaurus that was generated from the *Handbook of Artificial Intelligence* and *Collin's Dictionary*. The "Intelligent Intermediary for Information Retrieval" (I^3R), developed by Croft and Thompson (1987), consists of a group of "experts" that communicate via a common data structure, called the blackboard. The system consists of a user model builder, a query model builder, a thesaurus expert, a search expert (for suggesting statistics-based search strategies), a browser expert, and an explainer. Chen and Dhar (1991) incorporated a portion of the *Library of Congress Subject Headings* into the design of an intelligent retrieval system. The system adopted a branch-and-bound spreading activation algorithm to assist users in articulating their queries. 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, 1991; McCray & Hole, 1990). The UMLS includes a Metathesaurus, a Semantic Network, and an Information Sources Map. The Metathesaurus contains information about biomedical concepts and their representation in more than ten different vocabularies and thesauri.

Most of the knowledge bases adopted in these intelligent systems were either generated manually from domain experts, using the knowledge acquisition process, or derived from some existing thesauri (which were also created manually in the first place by some indexing/subject experts). A complementary approach to manual knowledge base creation is the *automatic thesaurus generation* approach.

Virtually all techniques for automatic thesaurus generation are based on the statistical co-occurrence of word types in text (Chen & Lynch, 1992; Crouch, 1990; Salton, 1989). Similarity coefficients are often obtained between pairs of distinct terms based on coincidences in term assignments to the documents of the collection. For example, a cosine computation can be used to generate normalized term similarities between 0 and 1. We discuss two similarity computations in detail in the next subsection. When pairwise similarities are obtained between all term pairs, an automatic term-classification process such as single-link or complete link classification can group into common classes all terms with sufficiently large pairwise similarities (Salton, 1978, 1989). The terms in the thesaurus classes can replace the initial search terms and be used to increase retrieval recall. Much of automatic thesaurus generation research has used sample document collections to generate term relationships and thesaurus terms have replaced search terms automatically without searchers' relevance feedback.

Despite research over the past two decades, there has been a lack of clear demonstration of the usefulness of using terms generated by co-occurrence analysis. Some research has shown that co-occurrence terms produce poor retrieval results when used in a fully automatic way (i.e., automatic query expansion) (Minker, Wilson, & Zimmerman, 1972; Peat & Willett, 1991; Smeaton & van Rijsbergen, 1983). However, recall improvements of the order of 10–20% have been demonstrated when the thesaurus is used in an environment similar to that in which the original thesaurus was constructed (Crouch, 1990; Salton, 1972; Salton & Lesk, 1971). Croft and Das (1990) discussed retrieval experiments in which users were prompted for additional terms (without a thesaurus, but using a searcher's own domain concepts) and reported significant improvements in retrieval effectiveness. The user-directed approach of selecting terms suggested by co-occurrence analysis data was also adopted in a recent experiment conducted by Ekmekcioglu, Robertson, and Willett (1992). They tested the retrieval performances of 110 queries on a database of 26,280 bibliographic (abstract) records using four approaches: original queries and query expansions using co-occurrence data, Soundex codes (a phonetic code that assigns the same code to words that sound the same), and string similarity measure (based on similar character microstructure), respectively. The four approaches produced 509 (original queries), 526 (term co-occurrence),

518 (Soundex), and 534 (string) documents, respectively. They concluded that there was no significant difference in retrieval effectiveness among these expansion methods and initial queries. However, a close examination of their results revealed that there was a very small degree of overlap between the retrieved relevant documents generated by the initial queries and those produced by the co-occurrence approach (19% overlap using the Dice coefficient). This suggests that search performance may be greatly improved, that is, a searcher can almost double the number of relevant documents retrieved, if a searcher can select and use the terms suggested by a co-occurrence thesaurus in addition to the terms he or she has generated. By allowing searchers to apply their domain knowledge and select terms from a co-occurrence thesaurus, search performance could be improved significantly.

After examining past research and its pitfalls closely, we believe that creating a robust and useful thesaurus automatically requires meeting the following conditions: (1) a complete document collection; (2) an appropriate co-occurrence function; and (3) interactive searcher relevance feedback (user-directed interaction). A thesaurus should represent the complete knowledge in the document collection and be kept up to date with its underlying database. That is, in order to create a useful thesaurus, the "source" of the knowledge should be as complete as possible. Only when a representative set of documents is acquired for a subject domain can we generate a thesaurus capable of providing clues to activate a searcher's domain concepts. Second, most prevailing co-occurrence analysis functions (e.g., cosine, Dice, Jaccard's) are symmetric in nature and have produced undesirable side-effects and poor performance (Peat & Willett, 1991). In this research we report an elaborate asymmetric co-occurrence analysis function, which was developed in an attempt to capture subject experts' domain knowledge and was based on extensive feedback from experts in several domains. More details will be provided below. Finally, an automatic thesaurus should also be used as a search aid and memory-jogging device during retrieval. The *automatic thesaurus generation* approach, if applied properly, can be extremely powerful in capturing the domain knowledge in textual databases and creating an environment for effective information retrieval.

The specific algorithms adopted in the research include: *term filtering, automatic indexing,* and *cluster analysis.* In the following section, we present an overview of these techniques.

- **Term Filtering:**
  Term filtering helps identify terms in a document that match with entries in existing domain-specific lexicons. Bates (1986) proposed a design model for subject access in online catalogs. She stressed the importance of building domain-specific lexicons for online retrieval purposes. A domain-specific, controlled list of keywords can help identify legitimate search vocabularies and help searchers "dock" on to the retrieval system. We used several domain-specific controlled lists of subject keywords, researchers' names, and organizational names for indexing in a Russian computing database (Chen & Lynch, 1992) (with about 200 MBs and 40,000 documents).

  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 a textbook), researchers' names (e.g., author indexes or researchers' directory), and other domain-specific objects (e.g., genes, experimental methods, organizational names, etc.) which are available online or can be obtained through OCR scanning. These domain-specific keywords can be used to help identify important concepts in documents automatically.

- **Automatic Indexing:**
  After term filtering the remaining texts may still contain many important concepts. Salton (1989) presents a blueprint for automatic indexing, which typically includes dictionary look-up, stop-wording, word stemming, and term-phrase formation. The algorithm first identifies individual words. Then, a stop word list is used to remove nonsemantic bearing words such as the, a, on, in, etc. After removing the stop words, a stemming algorithm is used to identify the word stem for the remaining words. Finally, term-phrase formation that formulates phrases by combining only adjacent words is performed.

- **Cluster Analysis:**
  Whereas *automatic indexing* identifies subject descriptors in a document, the relative importance of each descriptor to representing the content of the document may vary. Salton's *Vector Space Model* associates with each descriptor a weight to represent its descriptive power. Among the many probabilistic techniques that have been developed by various information science researchers, techniques which typically incorporate *term frequency* and *inverse document frequency* have been found to be simple and yet very useful (Salton, 1989). The basic rationales underlying these two measures are: terms which appear more times in a document should be assigned higher weights (*term frequency*), and terms which appear in fewer documents in the whole database (the more specific terms) should have higher weights (*inverse document frequency*).

  Based on *cluster analysis* (Everitt, 1980), the *Vector Space Model* has been extended for *automatic thesaurus generation* (or *automatic knowledge base generation*). The first stage in many cluster analyses is to covert the raw data (e.g., indexes and weights) into a matrix of interindividual *similarity, dissimilarity,* or *distance* measures. The result of a cluster analysis will be a number of groups, clusters, types or classes of individuals (Everitt, 1980). In *automatic thesaurus generation* (Chen & Lynch, 1993; Crouch, 1990), the most commonly used algorithms compute probabilities of indexes co-occurring in all documents of a database (sometimes referred to as *co-occurrence analysis*). Just

as a human inductive learning process generates concepts from a set of examples and benefits from the largest possible number of examples, a thesaurus created from a textual database becomes more "knowledgeable" as it becomes more subject-specific and larger in the size of its collections.

Although the above techniques had been employed in other applications, we have not seen them used together for developing a highly domain-specific, complete, and up-to-date automatic thesaurus for a community of scientific researchers. During our system development process, significant adaptation was required to meet the specific constraints and novel characteristics of the WCS.

An Application: The Worm Community System

The WCS's exhaustive collections and coverage of different communities' knowledge from formal to informal knowledge and from textual to graphical information have contributed significantly to its success in the first release. Among the functionalities of the WCS, online retrieval and browsing of rare and time-specific textual documents (especially the newsletter, the *Worm Breeder's Gazette,* dating back to 1975) has been considered one of its strengths. Formal and informal literature record in detail the progress and knowledge in the worm community.

In this research we used four main sources of textual documents in the WCS for thesaurus generation. Because these knowledge sources provided different forums for community members to discuss topics ranging from finished project and ongoing research, to laboratory observations and personal communications, the vocabularies used and concepts discussed tended to vary significantly. The four sources were:

- *The Worm Book:* This is the standard reference book, *The Nematode Caenorhabditis elegans* (Wood, 1988), used by community members. It contains 12 review chapters written by senior researchers on all aspects of *C. elegans* biology. It has about 700 pages of texts and figures. We generated an online version by using OCR software. We divided each chapter into 20–30 different *documents* based on its section and subsection structure. Each section and subsection that conveyed unique content and ideas was used as an individual document for analysis. In total, we generated 308 documents from the 12 chapters.
- *Journal abstracts:* One thousand four hundred sixty-seven (1467) *C. elegans* (refereed) articles with complete reference information and abstracts were acquired from *Medline* and *Biosis.* These documents occupied 1.6 MBs, with coverage from 1974 to 1992. On average, between 100 and 200 articles were published each year.
- *Worm Breeder's Gazette* (WBG): This is a newsletter, analogous to a moderated electronic bulletin board, which consists of short research items and (partial) results. The WBG has been published several times a year for over 10 years in an unrefereed, open format. Thanks to the assistance of worm community members, the WCS was able to acquire the complete collection of WBG since 1974, which comprised 1626 documents and required 4.6 MBs.
- *Conference proceedings:* The annual *C. elegans* meeting publishes one-page abstracts of presented papers which are cited as personal communications. Many ongoing research and preliminary results are revealed in these meetings. The WCS included 1313 documents from 1977 to 1992 that were 1 MB in size.

The four sources of knowledge comprised 4714 documents and 8 MBs of textual information. Sample entries are shown in Figure 1.

Identifying Descriptors: Term Filtering and Automatic Indexing

To identify candidate descriptors in each document, we performed term filtering and automatic indexing in order.

Term Filtering

We created the following lists of worm-related keywords with the help of several domain experts and from some external sources.

- *Researchers' names:* Researchers' names were extracted from the AUTHOR identifier of each document. By preprocessing all 4714 documents in WCS, we were able to identify 2709 unique author names. Any textual description in the abstract of a document which matched with these authors' names was identified as a researcher index.
- *Gene names:* One thousand five hundred twenty (1520) gene names were identified from the WCS (the WCS gene list) and the conference proceedings. (Among the four textual knowledge sources, only conference papers indicated what genes were involved in the research.) This worm-specific gene list was extremely useful for worm researchers because of the role the *C. elegans* worm plays in recent genetics research.
- *Experimental methods:* Thirty-seven (37) experimental methods were identified from the appendix of *The Worm Book.*
- *Subject descriptors:* Subject descriptors were created from two sources. We scanned all entries in the subject index of *The Worm Book* and we also incorporated a keyword list which was previously generated by worm researcher M. Edgley (former curator of *C. elegans* stock center). The total number of subject keywords in this list was 1048.

In total we identified 5314 worm-specific terms which can help identify important concepts in worm documents. Figure 2 shows sample entries in the four lists in alphabetical order.

Worm Book

```
LIT-TYPE: "WB"
REFERENCE: "W1.1"
TITLE: "I. GENERAL DESCRIPTION"
AUTHOR: "W. B. WOOD"
DATE: "1988"
ABSTRACT: "CAENORHABDITIS ELEGANS IS A SMALL, FREE-LIVING SOIL
NEMATODE FOUND COMMONLY IN MANY PARTS OF THE WORLD. IT FEEDS
PRIMARILY ON BACTERIA AND REPRODUCES WITH A LIFE CYCLE OF ABOUT
3 DAYS UNDER OPTIMAL
.....
THE SIMPLICITY, CONVENIENCE OF MANIPULATION, AND SHORT LIFE CYCLE OF
C. ELEGANS MAKE IT A USEFUL EXPERIMENTAL ORGANISM FOR THE STUDY OF
METAZOAN DEVELOPMENT AND BEHAVIOR."
```

Journal Abstracts

```
LIT-TYPE: "JOURNAL"
REFERENCE: "3"
AUTHOR: "ABDULKADER N;BRUIN JL"
TITLE: "INDUCTION, DETECTION AND ISOLATION OF TEMPERATURE-SENSITIVE
LETHAL AND/OR STERILE MUTANTS IN NEMATODES. 1. THE FREE-LIVING
NEMATODE CAENORHABDITIS ELEGANS"
DATE: "1978"
JOURNAL: "REV. NEMATOL"
ABSTRACT: "APPLYING A SERIES OF TECHNIQUES INTENDED TO INDUCE, DETECT
AND ISOLATE LETHAL AND/OR STERILE TEMPERATURE-SENSITIVE MUTANTS,
SPECIFIC TO THE SELF-FERTILIZING HERMAPHRODITE NEMATODE C. ELEGANS,
BERGERAC STRAIN (ABDULKADER ET BRUIN, 1976), 25 SUCH MUTANTS WERE
FOUND. OPTIMAL CONDITIONS FOR THE APPLICATION OF MUTAGENIC TREATMENT
AND THE DETECTION OF SUCH MUTATIONS ARE DISCUSSED."
SOURCE: "REV. NEMATOL 1978 1(1): 27-38"
```

FIG. 1. Sample knowledge sources.

Automatic Indexing

Automatic indexing was implemented mainly based on the procedure reported in Salton (1989). The following steps were executed in order:

- *Word identification:* Our system first identified words in each document, ignoring punctuation and case.
- *Stop-wording:* Next, we developed a "stop-word" list which included about 1000 common function (nonsemantic bearing) words such as on, in, at, this, there, etc., and "pure" verbs (words which are verbs only), for example, calculate, articulate, teach, listen, etc. This list was used to remove high-frequency words that were too general to be useful in representing document content.
- *Stemming:* We adopted a stemming algorithm to identify the word stem for each remaining word—a reverse of the *suffixing* procedure. The stemming component consisted of two parts. First, it included a 28,000-word (root words) dictionary with flags indicating legal suffixed forms. The total number of words our system recognized was about 80,000. We also included about 30 rules to interpret the flags for suffixes. The 22

Researchers (2,709 terms)	Genes (1,520 terms)
AAMODT E.	AAH/1
AAMODT E. J.	ABL-1
AAMODT ERIC	ABN
AAMODT ERIC J.	ACE-1
ABAD PIERRE	ACE-2
ABADON MONIQUE	ACE-3
ABDUL-KADER N.	ACR-1
ABDULKADER N.	ACT
...	...

Experimental Methods (37 terms)	Subject Descriptors (1,048 terms)
AXENIC GROWTH	A-BAND
BULK GROWTH ON BACTERIA	ACCESSORY STRUCTURE
CHEMOTAXIS	ACE
CLEAN	ACETYLCHOLINE
CROSS	ACETYLCHOLINESTERASE
EGG LAY	ACETYLCHOLINESTERASE INHIBITOR
FIXATE	ACT(ST15)
FREEZE	ACT(ST22)
...	...

FIG. 2. Sample term filters.

suffixes that could be stemmed by this algorithm included: ive, ion, tion, en, ions, ications, ens, th, ieth, ly, ing, ings, ed, est, er, ers, s, es, ies, ness, iness, and 's.

- *Term-phrase formation:* We then used the term-phrase formation technique to form phrases from adjacent words in the titles and abstracts of each document. After examining the subject descriptors typically used in *The Worm Book,* we decided to form phrases that contained up to three adjacent words—our system generated 1-word, 2-word, and 3-word phrases. For example, "DAUER," "LARVA," "FORMATION," "DAUER LARVA," "LARVA FORMATION," and "DAUER LARVA FORMATION" were generated from the three adjacent words "DAUER LARVA FORMATION" in a document. We will refer to these phrases simply as "terms" in the remainder of this article.

Automatic Worm Thesaurus Generation

After the concept descriptors for each document were identified we proceeded to perform term co-occurrence analysis for all documents in the WCS. A term weighting scheme based on the Vector Space Model (Salton, 1989) and an asymmetric similarity function (Chen & Lynch, 1992) were adopted for analysis. The blueprint for generating such a *concept space* (we refer to the thesaurus as a *concept space* to distinguish it from the *information space* represented by the WCS documents) is shown below:

- **Compute Term Frequency and Document Frequency:** We first computed the term frequency and the document frequency for each term in a document. Term frequency, tf_{ij}, represents the number of occurrences of term j in document i. Document frequency, df_j, represents the number of documents in a collection of n documents in which term j occurs. High term frequency indicates that a term is highly related to a document. High document frequency, on the other hand, indicates that a term is too general to be useful as a descriptor (i.e., no descriptive power).

  Usually terms identified from the title of a document are more descriptive than terms identified from the abstract of the document. In addition, terms identified by the term filters are usually more accurate than terms generated by automatic indexing. This is due to the fact that terms generated by automatic indexing are relatively "noisy." In our research, terms identified in titles were assigned heavier weights than terms in abstracts and terms identified by term filtering were assigned heavier weights than terms identified by automatic indexing.
- **Combine Weights:** We then computed the combined weight of term j in document i, d_{ij}, based on the product of "term frequency" and "inverse document frequency" as follows:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right)$$

where N represents the total number of documents in WCS and w_j represents the number of words in descriptor T_j. Multiple-word terms were assigned heavier weights than single-word terms because multiple-word terms usually conveyed more precise semantic meaning than single-word terms.

- **Perform Co-Occurrence Analysis:** We then generated a term co-occurrence table based on the asymmetric "cluster function" developed by the authors (Chen & Lynch, 1992). The limitation of the popular symmetric co-occurrence coefficients, for example, cosine, Dice, and Jaccard's, have been reported recently by Peat and Willett (1991). Their research showed that similar terms identified by symmetric co-occurrence functions tended to occur very frequently in the database being searched and thus did little or nothing to improve the discriminatory power of the original query. They concluded that this can help explain Sparck Jones' (1971) finding that the best retrieval results were obtained if only the less frequently occurring terms were clustered and if the more frequently occurring terms were left unclustered.

  We echo their observations and, in fact, we have independently reached the same conclusion through our experience in developing several thesauri for capturing subject experts' domain concepts (in terms of concepts and relationships) for several applications. In Chen and Lynch, (1992) we generated two knowledge bases automatically (one based on the cluster function and the other the popular cosine function (Everitt, 1980)) in a unique (former) East-bloc computing domain from a large textual database. We performed a memory-association experiment, comparing the recall and precision of the knowledge bases and four East-bloc computing experts in associating 50 randomly selected concepts (researchers, organizations, and subject keywords). The knowledge base that exhibited an asymmetric link property outperformed (statistically significant) the symmetric cosine knowledge base and all four human subjects in recalling relevant concepts in East-bloc computing. We believe the weighting-factor appearing in the equations below is an improvement of our asymmetric cluster algorithm. The asymmetric coefficient and weighting factor reward terms which are specific, that is, terms which improve the discriminatory power of the original query.

$$\text{Cluster weight } (T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times \text{weighting factor } (T_k)$$

$$\text{Cluster weight } (T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times \text{weighting factor } (T_j)$$

These two equations indicate the similarity weights from term T_j to term T_k (the first equation) and from term T_k to term T_j (the second equation). d_{ij} and d_{ik} were calculated based on the equation in the previous step. d_{ijk} represents the combined weight of both de-

scriptors T_j and T_k in document i. d_{ijk} is defined similarly as follows:

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right)$$

where tf_{ijk} represents the number of occurrences of both term j and term k in document i (the smaller number of occurrences between the terms was chosen). df_{jk} represents the number of documents (in a collection of N documents) in which terms j and k occur together. w_j represents the number of words of descriptor T_j. To *penalize* general terms (terms which appeared in many places) in the co-occurrence analysis, we developed the following weighting schemes:

$$\text{Weighting factor } (T_k) = \frac{\log \frac{N}{df_k}}{\log N}$$

$$\text{Weighting factor } (T_j) = \frac{\log \frac{N}{df_j}}{\log N}$$

Terms with a higher df_k value (more general terms) had a smaller weighting factor value, which caused the co-occurrence probability to become smaller. In effect, general terms were *pushed* down in the co-occurrence table (terms in the co-occurrence table were presented in reverse probabilistic order, with more relevant terms appearing first). This refinement was implemented after we tested our initial implementation with several biologists. They found that some very general (but not useful) terms, for example, gene, mutation, *C. elegans,* etc. were still suggested by the automatic thesaurus (at the top of the co-occurrence table). After imposing this penalty factor, the thesaurus was able to make more precise and specific suggestions.

Sample entries in the system-generated co-occurrence tables are shown in Figure 3. As shown in the co-occurrence table, "SPERM" was found to be most strongly related to researcher "WARD, S." (director of the Arizona worm laboratory who specializes in germline development and fertilization), with a weighted probability of 0.213660. In the second entry, "WORM COMMUNITY SYSTEM" was found to be most relevant to researcher "SCHATZ, B.," who is the director and architect of the WCS. His collaborators, including Powell and Peterson, also appeared near the top of this list. In the third and fourth entries, "CED-3" and "CED-4" were found to be relevant to "CED-9," which indicated that these terms were relevant to the topic of "PROGRAMMED CELL DEATH" in worm.

- **Apply Thresholds:**
  Without setting a probabilistic threshold for the co-occurrence table, the total number of co-occurrence pairs was 1,708,551. Each term may have a few thousand related concepts. This enormous size not only used a lot of memory, it may also overwhelm searchers during the thesaurus browsing process. For productive user–system interaction, only highly relevant concepts should be suggested to searchers.

1. WARD, S. : SPERM: 0.213660
2. WARD, S. : SPERMATOZOA: 0.185520
3. WARD, S. : MSP: 0.146670
4. WARD, S. : SPERMATOGENESIS: 0.142000
5. WARD, S. : PSEUDOPOD: 0.135020
6. WARD, S. : SPERMIOGENESIS: 0.131230

1. SCHATZ, B. : WORM COMMUNITY SYSTEM: 0.679110
2. SCHATZ, B. : WORM COMMUNITY: 0.679110
3. SCHATZ, B. : COMMUNITY SYSTEM: 0.679110
4. SCHATZ, B. : POWELL, K.: 0.679070
5. SCHATZ, B. : PETERSON, L.: 0.679070
6. SCHATZ, B. : SOFTWARE: 0.652560

1. CED-9 : CED-4: 0.301500
2. CED-9 : CED-3: 0.289330
3. CED-9 : CELL DEATH: 0.275170
4. CED-9 : CED-9(LF): 0.262330
5. CED-9 : DEATH: 0.244440
6. CED-9 : PROGRAMMED CELL: 0.234000

⋮

1. CELL DEATH : DEATH: 0.434510
2. CELL DEATH : PROGRAMMED CELL DEATH: 0.368510
3. CELL DEATH : CED-3: 0.262800
4. CELL DEATH : PROGRAMMED: 0.213700
5. CELL DEATH : CED-4: 0.206270
6. CELL DEATH : PROGRAMMED CELL: 0.201770

FIG. 3. Sample co-occurrence table.

After consulting worm biologists in the Arizona worm laboratory to decide on a reasonable number of related terms for each concept, we chose 100 as the maximum number of links for any node. This effectively removed about 60% of the less relevant co-occurrence pairs. The resulting worm thesaurus contained 709,659 pairs of related concepts. After applying the thresholds, the total number of unique terms found in the four sources was 7829. On average, each term had about 90 relevant concepts.

- **Automatic Thesaurus Generation Using Combined Sources versus Multiple Sources:**
  As stated earlier, WCS contained knowledge sources of different natures, some more formal than others. The vocabularies and topics discussed in each of the knowledge sources also varied widely. To investigate the effect of "knowledge discovery" using separate sources versus one combined source, we created two versions of the worm thesaurus. One version, called kb1, treated the four WCS knowledge sources as one, that is, the size of the database was 4714 (documents). Co-occurrence analysis was performed only once for the entire collection. Conversely, kb2 analyzed each knowledge source individually. We first ran co-occurrence analysis against all documents in *The Worm Book.* We then ran the same procedure for journal articles, *Worm Breeder's Gazette,* and proceedings articles, respectively. Results from the four separate thesaurus generation processes were then combined to form one thesaurus. Table 1 shows the number of nodes found in each knowledge source, and Table 2 shows the number of links found in each knowledge source.

TABLE 1. Number of terms found in each knowledge source.

Concept Types	Worm Book	Journal Articles	WBG	Conference Proceedings	Total
Researchers	12	1356	1343	1543	2709
Genes	277	20	740	406	798
Experimental Methods	14	15	17	9	20
Subject Descriptors	1044	1673	2526	1339	4302
Total	1347	3064	4626	3297	7829

We hope that some interesting patterns may be discovered when analysis is performed on individual sources. In theory, kb2 thesaurus generation should also be faster because of the nature of the cluster function (an $O(n^2)$ algorithm). Partitioning the entire WCS database into four sources makes the number of descriptors, n, significantly smaller, which helps reduce the CPU times required for the thesaurus generation process.

Thesaurus Evaluation: An Experiment

Our prototype system was developed in ANSI C and ran on both Sun Sparc stations and DECstations. It took 9.2 and 4 hours of CPU time to generate kb1 and kb2, respectively. The resulting sizes for kb1 and kb2 were 12.3 MBs and 12.6 MBs, respectively.

We also developed an X-Windows interface to use the thesaurus. Sample interactions are shown in Figures 4 and 5. In Figure 4, terms relevant to "neurons," "sensilla," "dauer larva," and "chemotaxis" (entered by user) are displayed in weighted order. Users can select among these system-suggested terms to activate the thesaurus again (user-selected terms are highlighted with markers), or they can enter their own terms to activate the thesaurus (as shown in the pop-up window). Terms selected by a user during the iterative thesaurus browsing process are recorded in a separate area as shown in Figure 5.

To evaluate the performance of the two thesauri and to pinpoint directions for improvement, we conducted an experiment in Winter 1992 with six subjects who were affiliated with the Arizona Worm Laboratory. These subjects included two experts, two novices, and two outsiders. The research questions we aimed to address in the experiment were stated as follows:

(1) Will the thesaurus be able to help find relevant terms and improve (concept) recall and precision for searchers of different backgrounds?
(2) Which thesaurus is better—kb1 or kb2? Are relevant terms displayed in the order of relevancy, beginning with more relevant terms?
(3) What, if any, are the novel characteristics of the system-generated thesauri? What are the problems and what improvements are needed?

These questions must be answered before either thesaurus is implemented and integrated into the existing Worm Community System.

TABLE 2. Number of links found in knowledge source.

Link Types	Worm Book	Journal Articles	WBG	Conference Proceedings	Total
Author-author	4	5,526	5,666	6,078	17,274
Author-gene	541	57	8,823	2,640	12,061
Author-method	62	383	1,303	336	2,084
Author-subject	3,578	36,965	70,400	30,465	141,408
Author related	4,185	42,931	86,192	39,519	172,827
Gene-author	541	57	8,820	2,640	12,058
Gene-gene	5,076	0	22,334	3,558	30,968
Gene-method	405	0	1,491	305	2,201
Gene-subject	22,129	439	89,538	18,528	130,634
Gene related	28,151	496	122,183	25,031	175,861
Method-author	62	383	1,287	336	2,068
Method-gene	405	0	1,490	305	2,200
Method-method	64	44	100	20	228
Method-subject	2,733	2,713	7,627	1,666	14,739
Method related	3,264	3,140	10,504	2,327	19,235
Subject-author	3,578	36,256	68,558	30,364	138,756
Subject-gene	22,129	439	89,365	18,528	130,461
Subject-method	2,733	2,712	7,671	1,666	14,782
Subject-subject	162,892	226,863	545,249	121,625	1,056,629
Subject related	191,332	266,270	710,843	172,183	1,340,628
Total	226,932	312,837	929,722	239,060	1,708,551

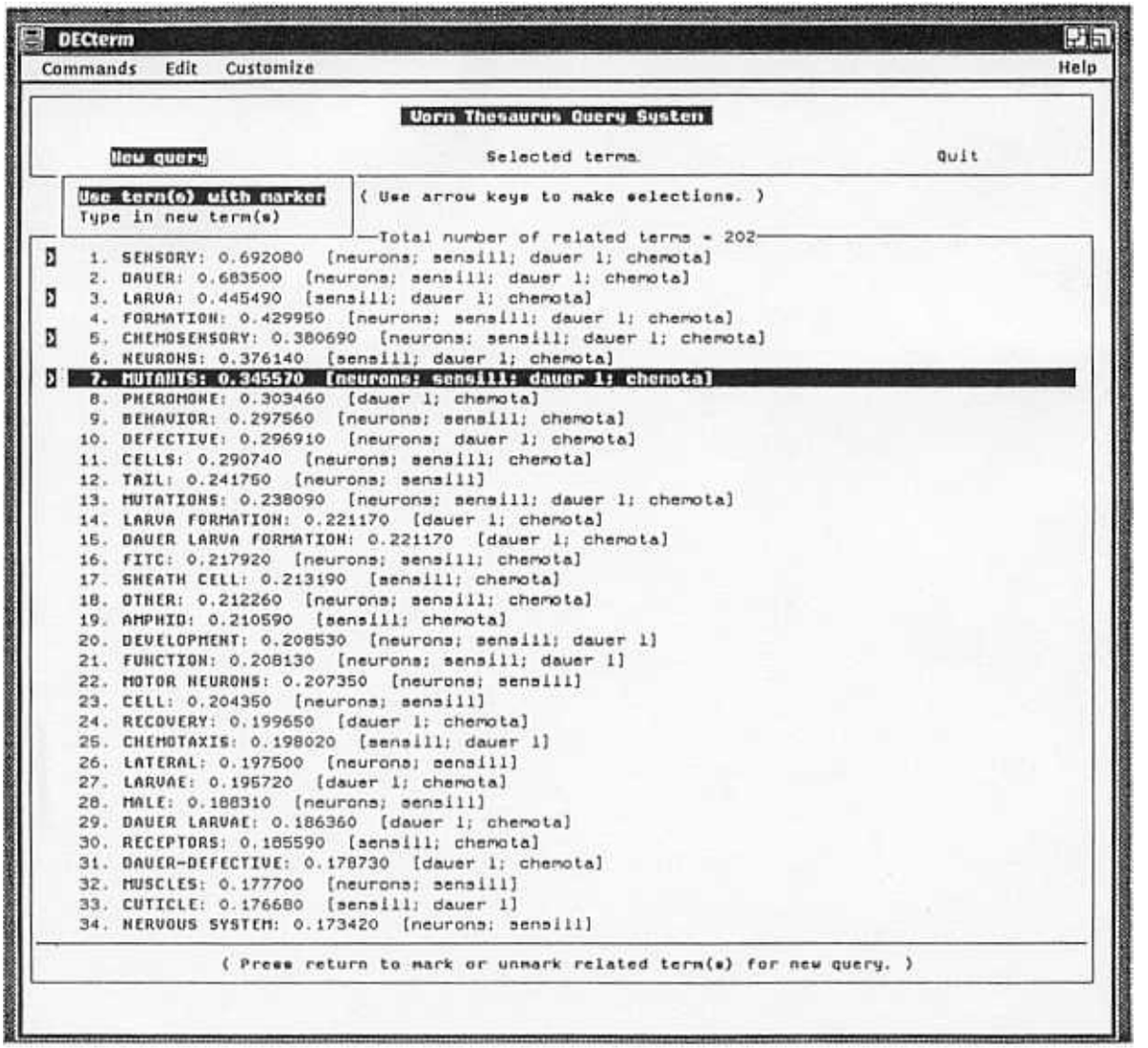

FIG. 4. Thesaurus terms relevant to neurons, sensilla, dauer larva, and chemotaxis.

Experimental Design

The experiment consisted of two parts: a term association experiment and a searcher browsing experiment. Six subjects with different backgrounds were selected to investigate the effects of a searcher's expertise on thesaurus usage. Subjects 1 and 2 were considered experts in molecular and cellular biology (MCB). They have worked in the Arizona Worm Lab for several years and have published papers in this area. Subject 1 was the lab manager and subject 2 was an advanced-stage Ph.D. student. Subjects 3 and 4 were considered novices; one was a master's student in MCB and the other an undergraduate senior. Both worked in the worm lab. Subjects 5 and 6 were worm outsiders. Subject 5 was a new Ph.D. student in MCB with an master's degree in computer science and had working experience in fly biology. Subject 6 was a graduate student in ecology and evolutionary biology at the University of Arizona. His research area was mainly in the evolution of animal size and lifespan.

- **Term Association Experiment:**
  The first step of the term association experiment was to give each subject a preselected term. Sixteen terms chosen with the help of several worm researchers were presented to each subject in order. Terms included researchers' names, gene names, and subject descriptors. The subjects were asked to write down concepts (genes, researchers, methods, and subject descriptors) related to each preselected term. A sample experimental sheet for the first term, "WARD, SAMUEL" (director of the Arizona Worm Lab), and the 24 related terms generated by subject 1 are shown in Figure 6.

  Subjects were then asked to mark terms suggested by kb1 and kb2 as irrelevant, somewhat relevant, or very relevant. kb1 and kb2 were presented to the subjects in random order and all terms were listed in de-

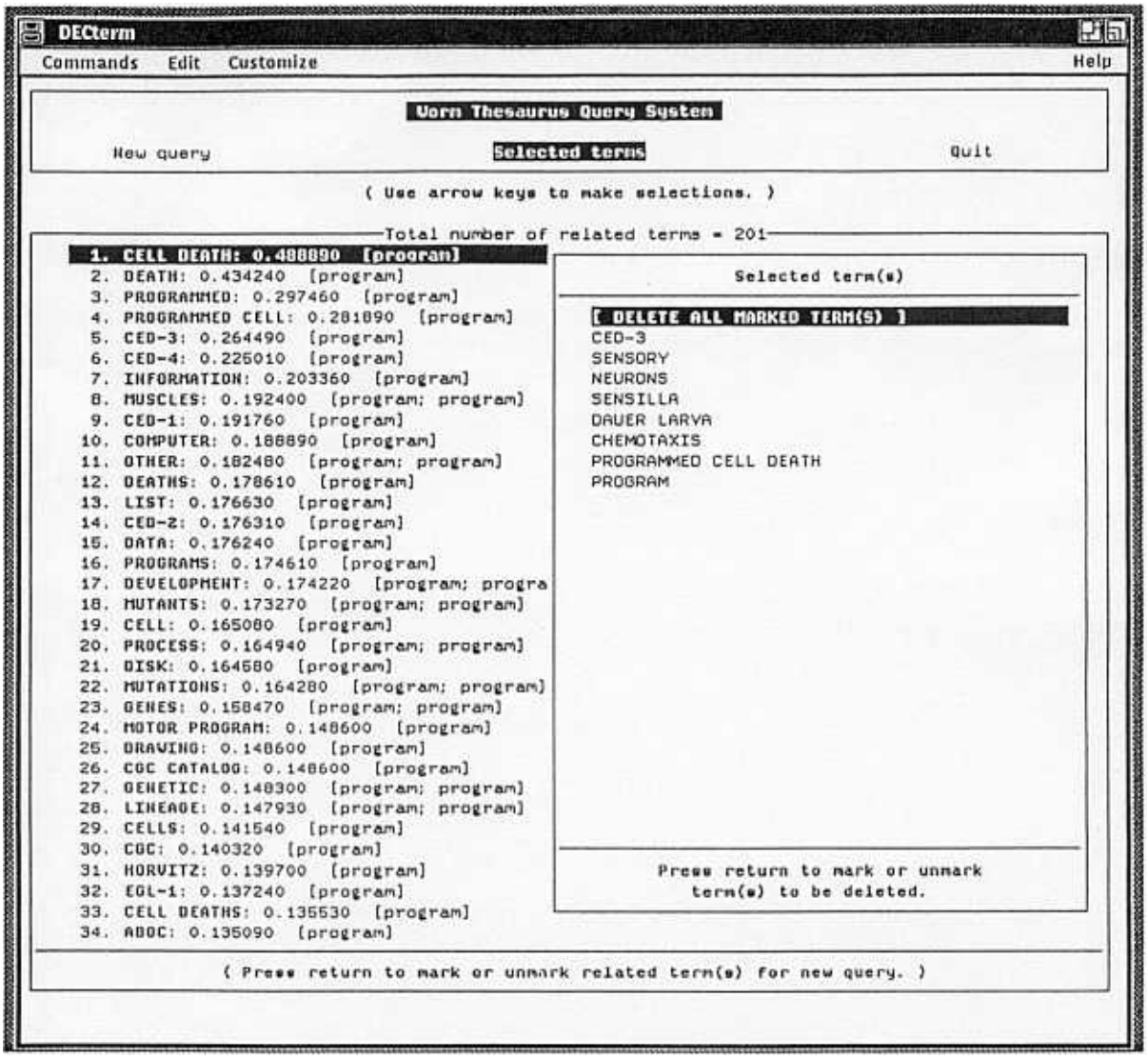

FIG. 5. Terms selected by the user during thesaurus browsing.

creasing order of relevance on a single page (48 related terms in total were displayed on each page for any given term). After selecting system-suggested terms, subjects were asked to evaluate their own terms again and they were allowed to remove terms which they no longer considered relevant.

- **Searcher Browsing Experiment:**
  After the term-association experiment, subjects were

PLEASE WRITE AS MANY TERMS OR CONCEPTS AS POSSIBLE THAT
RELATE TO THE FOLLOWING TERM. INCLUDE AUTHORS, GENES,
METHODS AND SUBJECTS THAT ARE RELEVANT.

1) WARD, SAMUEL

SPERMATOGENESIS	SPE GENES
FER GENES	MSP (MAJOR SPERM PROTEIN)
GENOME EVOLUTION	MOTILITY
ASSEMBLY	DIFFERENTIATION
MORPHOGENESIS	WORM COMMUNITY SYSTEM
SPE-26	SPE-12
FER-1	SSP (SPERM SPECIFIC PRODUCT)
CYTOSKELETON	AGING/LONGEVITY
CHEMOTAXIS	NEUROANATOMY
L' HERNAULT	SHAKES
MIWA	VARKEY
ELECTRON MICROSCOPY	SEQUENCING

FIG. 6. Subject-suggested descriptors.

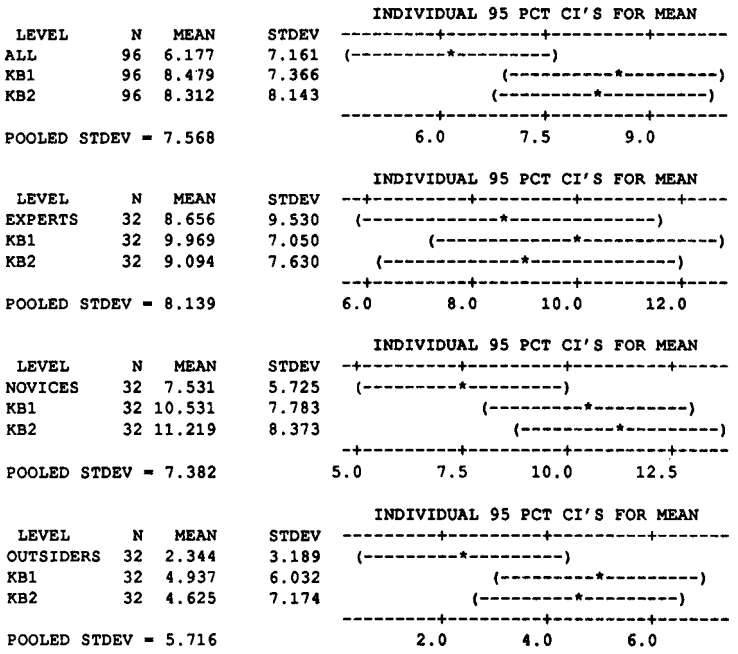

FIG. 7. ANOVA analysis for relevant terms.

asked to browse the online worm thesaurus freely—using any terms they preferred and exploring any way they liked. Most subjects used terms that they were either interested in or were familiar with. This experiment gave us a better idea of how the thesaurus could be used for more real-life purposes and by different user groups. The online thesaurus suggested additional terms, which were shown on an X-Windows display. During browsing subjects were asked to think aloud and gave specific comments, observations, or suggestions. The subjects browsed both thesauri simultaneously and compared the results. Their complete online sessions were logged and verbal protocols were recorded and transcribed for analysis. The complete experiment lasted between 1.5 hours and 2.5 hours for each subject.

Analysis and Experimental Results

We summarize results pertaining to the three research questions below.

- *Finding more relevant terms:* By counting the numbers of terms generated by the subjects themselves and the system-suggested terms marked relevant (both somewhat relevant and very relevant) by the subjects, we were able to tabulate and analyze whether the two thesauri were able to contribute relevant terms during a retrieval process. An analysis of variance procedure (ANOVA) using the MINITAB statistical package (Ryan, Joiner, & Ryan, 1985) was conducted for the search terms, followed by a two-sample *t*-test to determine the differences in means. The results are summarized in Figure 7. Overall, for each term kb1 was able to suggest 8.479 terms and kb2 suggested 8.312 terms. Subjects were able to generate 6.177 terms by themselves. Not surprisingly, the expert group performed better than the novice group, which performed better than the outsider group in generating relevant terms.

  For kb1, the two-sample *t*-test revealed that there were statistically significant differences (at 10% confidence interval level) in means for (ALL vs. KB1; p = .029), (NOVICES vs. KB1; p = .084), and (OUTSIDERS vs. KB1; p = .037). For kb2, there were statistically significant differences (at 10% confidence interval level) in means for (ALL vs. KB2; p = .055) and (NOVICES vs. KB1; p = .084). There were no statistical differences between the numbers of terms suggested by kb1 and kb2. In conclusion, kb1 and kb2 performed equally well and they helped identify more relevant terms for subjects, especially for novices and outsiders (there was a very small degree of overlap between a subject's terms and those suggested by the kb).
- *Concept recall and concept precision.* In contrast to the *document* recall and precision measures typically used in information science research, we adopted *concept recall* and *concept precision* for evaluation. Instead of examining the number of relevant documents retrieved, we counted the number of relevant terms (concepts) identified by the thesauri. These two measures were considered appropriate for evaluating the quality of term-association in thesauri (Chen & Lynch, 1992). They were computed as follows:

$$\text{Concept recall} = \frac{\text{Number of retrieved relevant concepts}}{\text{Number of total relevant concepts}}$$

$$\text{Concept precision} = \frac{\text{Number of retrieved relevant concepts}}{\text{Number of total retrieved concepts}}$$

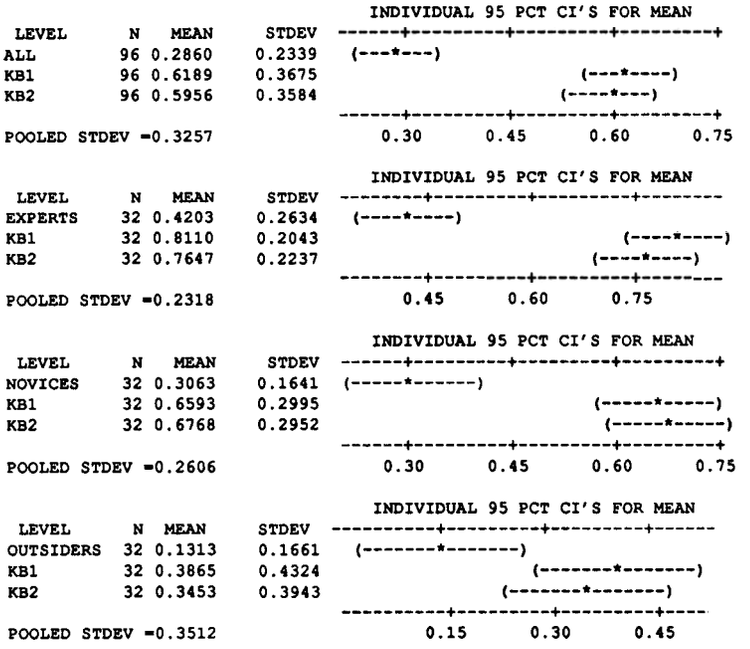

FIG. 8. ANOVA analysis for concept recall.

For all subjects, the terms they initially generated and the terms selected from kb1 and kb2 were included to represent the *total relevant concepts*—the target set of concepts which can be obtained through user–thesaurus interaction. Based on this target set of concepts, we can examine the subjects' initial terms (generated without any thesaurus help) and determine the subjects' *concept recall* and *concept precision* levels when the thesaurus component was unavailable, that is, by counting the number of terms that matched the target terms. We then evaluated the *concept recall* and *concept precision* levels for kb1 and kb2 by counting the number of thesaurus terms which matched with the target terms. Both ANOVA tests and two-sample *t*-tests were performed for *concept recall* and *concept precision.*

The ANOVA results for *concept recall* are shown in Figure 8. Overall, there were significant differences among the subject groups, kb1, and kb2. kb1 and kb2 were significantly better than all human subjects in recalling concepts. On average, subjects' recall level was at 28.60%; whereas kb1 and kb2 were at 61.89% and 59.56%, respectively. For experts, novices, and outsiders, the two thesauri consistently outperformed them in recalling concepts (all two-sample *t*-tests revealed statistically significant differences at the 10% level). This was to be expected, however, considering the fact that the two thesauri analyzed all documents in the complete WCS to generate a complete and up-to-date *concept space* for worm biology. The thesauri were exhaustive and can become an *organizational memory* for the entire worm community. Human subjects (even experts), on the other hand, were severely constrained and hindered by their experience, expertise, and the cognitive demand for recalling long-term memory in a short period of time. A thesaurus could potentially be very useful as a memory-jogging and concept-exploration tool for searchers of all levels.

As shown in Figure 9, the thesauri produced low precision levels, compared with those produced by the human subjects. Human subjects had about a 77.08% concept precision level; kb1 and kb2 had 24.17% and 23.66% precision levels, respectively. The low precision levels of thesauri were due to the noise terms (general terms for the most part) in the thesauri and the subjective determination of the relevancy of those terms to the subjects' queries and needs. As is evident in information science research, even man-made thesauri are only useful when presented in the context of the searchers' needs and when selected by the searchers themselves. Thesauri should be used for *consultation* purposes, not for automatic term replacement. Searchers' involvement during the thesaurus consultation process is crucial to the success of thesaurus usage.

In conclusion, the thesauri appeared to be better at recalling relevant concepts than human subjects; but human subjects were more precise than the thesauri. kb1 and kb2 achieved similar levels of performance in both recall and precision. With joint human–computer collaboration, it appears that an automatic thesaurus-augmented search process can become very fruitful and productive.

- *Subjective evaluation: Problems and novel characteristics.* After the browsing experiment, the subjects' verbal protocols were collected and analyzed. The protocols represented a subjective and qualitative evaluation of how each subject felt about the two thesauri when using them.

  Many problems pointed out by the subjects were related to our automatic indexing procedure, which included stemming, stop-wording, and term-phrase formation. The main problems reported included:

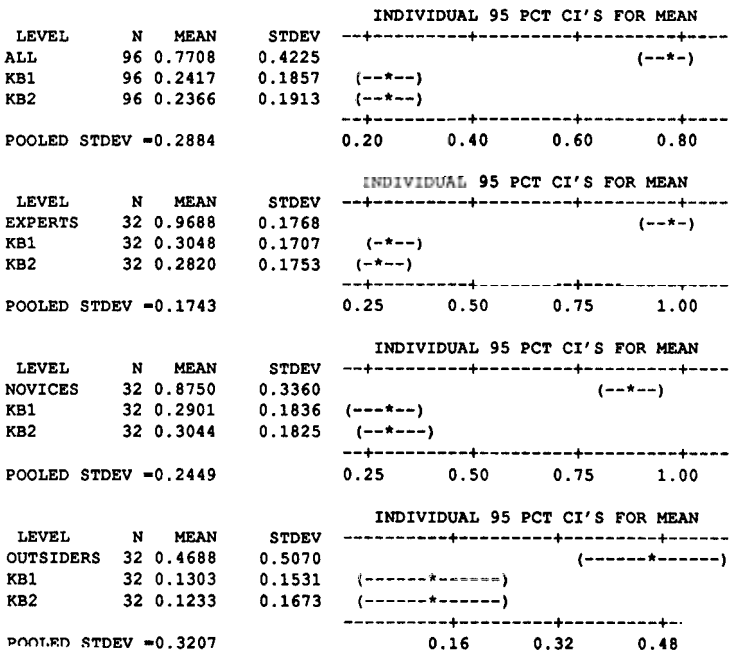

FIG. 9. ANOVA analysis for concept precision.

- Some words were misspelled, for example, AGE-] should be AGE-1 (error due to OCR scanning) or words were improperly stemmed, for example, C. ELEGAN should be C. ELEGANS.
- Often authors appeared more than once in the thesaurus, but in different forms, for example, JOHN, T. and JOHN, TOM.
- Some authors who were very relevant to a term were missing, for example, "I am surprised it didn't pull up Jonathan Hodgkin (in relation to MATING) . . . seems it is missing some of the people." This was due to the threshold we adopted, which removed some relationships by accident.
- Words that normally appear together were separated by the system, for example, "VAS . . . should appear with DEFERENS. You wouldn't find VAS by itself."
- Many terms suggested by the system were either irrelevant or too general to be applicable to a search term, for example, "FACT is such a generic term. . . ."
- The first terms suggested by the system were sometimes exact synonyms for the search term, for example, "See, now the first four here [POINTING TO kb1], five here . . . are worthless . . . because they are . . . identical to the [search] word." Even though reported as a problem, the system's ability to bring up synonymous terms may be quite useful for online search, especially when the initial search terms have failed to generate any hits.

Despite the above problems, most subjects found the thesauri to be interesting and useful, especially as a memory-jogging device or as a learning tool. Some important observations made by the subjects included:

- Overall, the system is useful and good, for example, "It looks like you are in the right direction . . . it looks good." (subject 1—S1), "Cool . . ." (S4), "It's pretty good . . . I think that's very helpful." (S5); "That seems like a lot of relevant terms for searching." (S6). Except for one novice subject who was more computer-resistant (she expressed her uneasiness with computer use), the subjects found the thesauri useful for various reasons.
- For the most part, more relevant terms appeared to have been listed first, for example, "It is getting more general down here." [POINTING TO BOTTOM OF LIST] (S1). "The first eleven are great, but after that, they're not so good . . . the first five are . . . quite relevant, . . . 6 and 7 are way too general . . . I'd say it's very good initially . . . and then . . . there's a fairly low frequency of relevance." (S5). "It is getting more general down here. . . . Is this deliberate? The way that more general terms are down here?" (S6). However, for some terms it was not so obvious. Comments reflecting interspersion of relevant terms were: "Some terms on the bottom of the list are still good like this one." (S1); ". . . DIVISION PATTERN is very relevant . . . which is number 18." (S5). The thesaurus' ability to list more relevant terms first is important, especially from the perspective of designing an effective and precise thesaurus browsing interface.
- Learning, serendipitous browsing, and memory-jogging occurred frequently. Many subjects found something "interesting" or unexpected that would help them in their queries. This is particularly obvious for novices and outsiders, who were often amazed by the thesaurus's ability to relate genes,

researchers, and subject topics. We postulate that when an expert begins to explore an unknown research topic (thus becoming a novice), the same learning process aided by the thesaurus could occur. For all subjects, the thesaurus also served as an excellent tool to remind them of something they previously had forgotten. Some sample comments related to these observations include: ". . . this is doing very well . . . I mean extremely well . . . this has a whole bunch of things that I'm very interested in. . . . This is potentially very useful. . . to me, I mean, I'm the novice . . ." (S5). [Using LONGEVITY in search] ". . . I am not familiar with people on this. A lot of these are. . . probably people who do stuff about this. . . pumping. . . I have no idea. There might be some paper. Oh yeah, spermgio pumping. . . . When you first see pumping is like what and then you realize it is from spermgio pumping. . ." (S6).

In summary, the results from the experiment were very encouraging. Both kbs suggested relevant terms and concepts that would not only be helpful for different users, but useful in spurring user ideas and desire to acquire knowledge. Both kb1 and kb2 were tested by objective and subjective measures and produced results indicating that they did produce relevant terms, improve concept recall, and would be useful to the Worm Community System.

Inconclusive was evidence of whether kb1 or kb2 suggested more relevant terms. The problem of discernment here was that both produced many of the same relevant terms and even if the terms were different, they were useful for different reasons. Some subjects commented that kb1 appeared to suggest more relevant genes, whereas kb2 suggested more relevant authors. However, this observation was not consistent among all subjects.

Conclusions and Future Directions

Since the experiment, we have made several changes and have fine-tuned our algorithms according to the subjects' suggestions. We have removed the stemming procedure from automatic indexing to avoid creating noise and ungrammatical phrases, for example, CLONING will not be stemmed as CLONE (one is a process, the other is the output), C. elegans will not be stemmed to C. elegan which is ungrammatical, etc. Despite removing the stemming routine, the number of unique descriptors identified from all the documents did not change significantly. We created a domain-specific stop-word list for worm biology which contained about 600 very general molecular biology terms such as gene, process, mutation, etc. This list helped us remove many general (and thus irrelevant) terms in the thesaurus. We standardized all researchers' names according to the format of last name, followed by the first character of the first name. This helped remove the problem of same names appearing in different forms. We also included allele for gene names to convey more precise meaning. This was a suggestion made by one expert subject. Because a gene and a gene with allele have different meanings, for example, daf-9 and daf-9(e1406), we modified the text processing routine to identify daf-9(e1406) as well. Gene names with allele are now captured in the worm thesaurus.

We have incorporated one version of the thesaurus into the WCS (kb2). An X-Windows thesaurus-browsing interface which can accept multiple terms, identify other relevant terms by means of the thesaurus, combine the weights associated with terms, and rank terms in order, has been developed, as shown in Figure 10. Searchers can use it to elicit suggested terms from the thesaurus. In the recent *C. elegans* meeting held at the University of Wisconsin at Madison in June 1993 (with about 650 participants), the WCS and the worm thesaurus were used extensively and perceived favorably by the worm biologists during the three-day live demo sessions. They were able to consult the worm thesaurus whenever they wished to explore other topics relevant to their initial queries. Many graduate students and post doctoral researchers, in particular, expressed great interest in the WCS and the thesaurus, especially for their potential value to help them quickly enter a new and dynamic research area. Many viewed the WCS as a comprehensive, tightly integrated *electronic library* with an online reference librarian (i.e., the worm thesaurus and its browsing interface).

As a long-term effort to develop a more efficient and "intelligent" framework and design for the management, retrieval, sharing, and dissemination of information for distributed, scientific computing, our planned research directions include the following:

- *Incremental thesaurus generation.* Currently, the thesaurus was generated in a batch mode (in about 4 hours). As the WCS database becomes larger, an effective method for incremental update for the thesaurus will be needed. We are currently developing an incremental version of our cluster algorithm. Storing some intermediate results for term frequencies and inverse document frequencies should make an incremental update for the thesaurus possible.
- *Handle larger applications.* This research illustrated the feasibility of an automatic, domain-independent approach to the creation of online thesauri for scientific domains. We are currently exploring other research communities, which may possess even larger amounts of information and which experience the same *information overload* and *vocabulary* problems in information retrieval (e.g., fly database, human genome database, physics literature database, etc.). By testing our approach in more communities, we will be able to verify and fine-tune our framework and research techniques.
- *Capture the fluidity of concepts.* So far we have not included the "time" dimension of the documents and concepts in our analysis. By time-tagging each concept and analyzing the activities associated with it (e.g.,

Concept Space Thesaurus in Worm Community System release 2

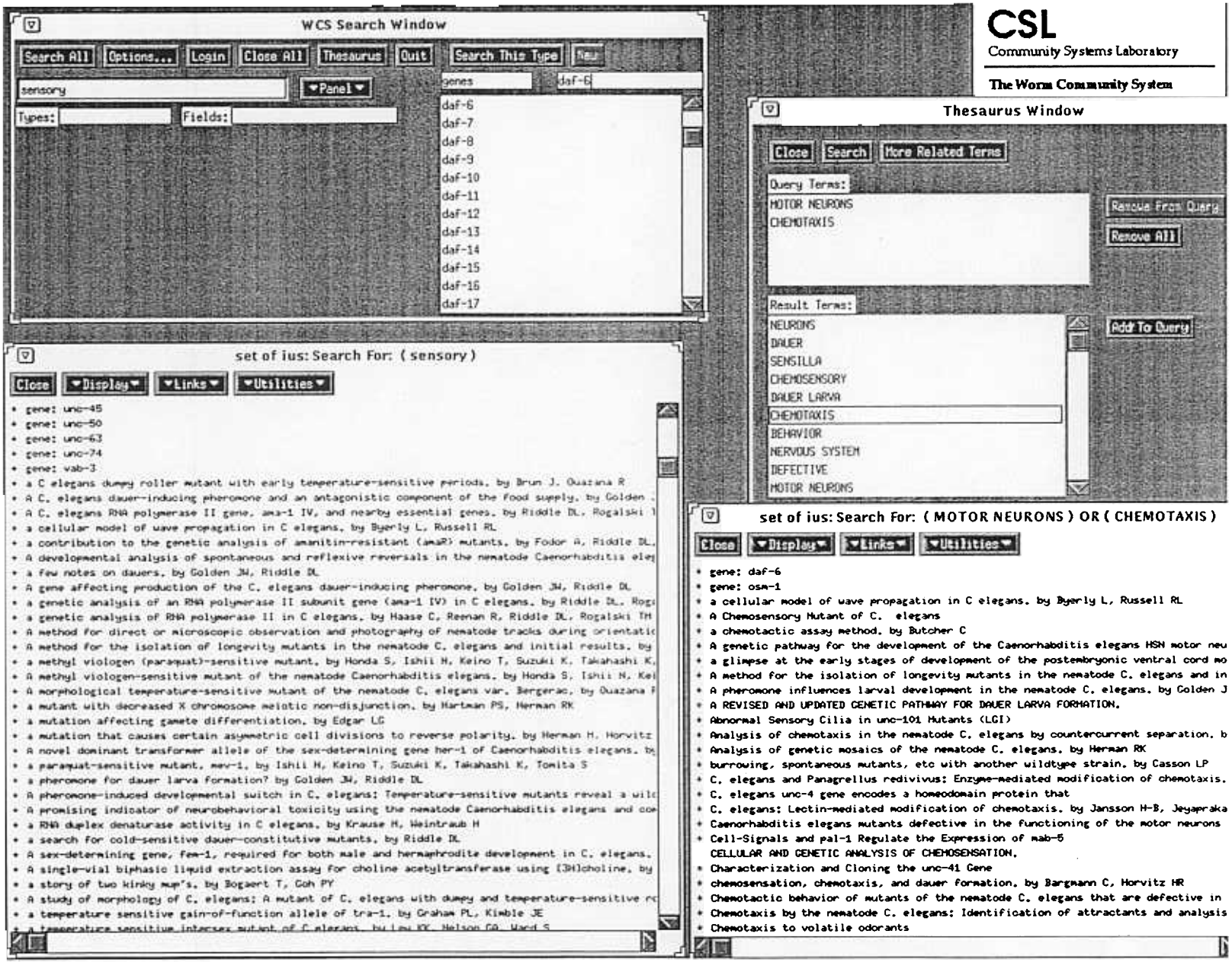

FIG. 10. Concept thesaurus in the Worm Community System.

when it first appeared, when it was most actively used, etc.), we believe a more fluid and time-precise thesaurus can be created.

- *Automatic thesaurus consultation.* As a natural extension of the current research, we will be testing some AI-based general search algorithms (e.g., branch-and-bound and Hopfield network) (Chen et al., 1993) for automatic thesaurus consultation. These algorithms will be able to assist searchers in traversing the entire *concept space* by following the more relevant links first, a general characteristic of optimal or heuristic search algorithms. We have done some work in this area already, but significant experimentation is still required to develop a robust automatic thesaurus consultation module for the WCS.
- *A concept space approach to solving the vocabulary problem.* Finally, we believe we are moving closer to finding a solution to one of the most challenging problems in information retrieval—the *vocabulary* problem. In scientific communities an outsider (e.g., a fly biologist) often needs to search for literature in other domains (e.g., worm biology) using his/her own vocabularies (i.e., fly-specific terms). By adopting our approach, we can create another concept space, say, for the fly community, by acquiring and analyzing a good collection of fly documents. By identifying the overlapping terms (i.e., common biology concepts) in the two concept spaces—the searcher's fly concept space and the target database's worm concept space—and by adopting a multiple-thesauri consultation process (we have developed one in Chen et al. [1993]), a searcher's fly-specific terms can be used to traverse the two concept spaces and eventually converge toward specific terms in the (target) worm concept space. We already have developed a fly thesaurus (Chen et al., 1994) and we are in the process of experimenting with determining how to intersect and traverse the fly-worm concept spaces.

  The rationale behind our approach is that, instead of requiring knowledgeable information specialists (knowledgeable in several subject areas) to perform term matching and consultation, we could automatically create different domain-specific thesauri tailored to the vocabularies and concepts exhibited in related disciplines. For example, biology disciplines might be classified based on: (1) *organisms,* coli, yeast, worm, fly, mice, and human; or (2) *level of analysis,* biochemical, molecular, cellular, developmental, medical, and ecological. Each of these classifications has a specific literature, which can be used to generate a domain specific thesaurus automatically. Our longer-term research effort will involve creating different "outsider" thesauri and incorporating them into the Worm Community System. We hope by expanding the "knowledge" of the WCS, the community system will eventually be able to assist in seamless, distributed, and "intelligent" concept-based information retrieval for different users from other communities.

Acknowledgments

This project was supported mainly by two NSF grants: the NSF CISE Research Initiation Award, IRI-9211418, 1992–1994 (H. Chen, "Building a Concept Space for an Electronic Community System") and the NSF CISE Special Initiative on Coordination Theory and Collaboration Technology, IRI-9015407, 1990–1993 (B. Schatz et al., "Building a National Collaboratory Testbed"). We wish to thank Ed Grossman and Terry Friedman for implementing WCS and Kevin Powell for helping to integrate the thesaurus into WCS. We would also like to thank the faculty and students of the Molecular and Cellular Biology Department at the University of Arizona for their kind assistance and valuable suggestions, in particular, those of Samuel Ward, Bill Achanzar, John Calley, Alicia Minniti, Paul Muhlrad, Wayne Van Voorhies, and Andrea Wellington.

References

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

Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information Processing and Management, 22,* 229–241.

Carmel, E., Crawford, S., & Chen, H. (1992). Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man and Cybernetics, 22,* 865–884.

Chen, H., & Dhar, V. (1990). User misconceptions of online information retrieval systems. *International Journal of Man-Machine Studies, 32,* 673–692.

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management, 27,* 405–432.

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

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

Chen, H., Schatz, B., Martinez, J., & Ng, D. T. (1994). *Generating a domain-specific thesaurus automatically: An experiment on FlyBase.* (Working Paper, CMI-WPS 94-02.) Center for Management of Information, College of Business and Public Administration, University of Arizona.

Courteau, J. (1991, Oct. 11). Genome databases. *Science, 254,* 201–207.

Croft, W. B., & Das, R. (1990, Sept.). Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the 13th Conference on Research and Development in Information Retrieval.* Brussels, Belgium.

Croft, W. B., & Thompson, R. H. (1987). *I³R*: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science, 38,* 389–404.

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

Ekmekcioglu, F. C., Robertson, A. M., & Willett, P. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Science, 18,* 139–147.

Everitt, B. (1980). *Cluster Analysis* (2nd ed.). London: Heinemann.

Foss, C. L. (1989). Tools for reading and browsing hypertext. *Information Processing and Management, 25,* 407–418.

Fox, E. A. (1987). Development of the CODER system: A testbed for artificial intelligence methods in information retrieval. *Information Processing and Management, 23,* 341–366.

Frenkel, K. A. (1991). The human genome project and informatics. *Communications of the ACM, 34,* 41–51.

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,* 964–971.

Lander, E. S., Langridge, E., & Saccocio, D. M. (1991). Mapping and interpreting biological information. *Communications of the ACM, 34,* 33–39.

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.* 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.* Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Minker, J., Wilson, G. A., & Zimmerman, B. H. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval, 8,* 329–348.

Monarch, I., & Carbonell, J. G. (1987, Spring). CoalSORT: A knowledge-based interface. *IEEE EXPERT,* 39–53.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42,* 378–383.

Piatetsky-Shapiro, G. (1989). Workshop on knowledge discovery in real databases. In *International Joint Conference of Artificial Intelligence.* Detroit, MI: Morgan Kaufmann Publishers, Inc.

Pool, R. (1993, Aug. 13). Beyond database and e-mail. *Science, 261,* 841–843.

Quint, M. (1991). Banks looking more closely at their credit card holders. *New York Times,* May 27, p. 1.

Ryan, B. F., Joiner, B. L., & Ryan, T. A. (1985). *MINITAB handbook* (2nd ed.). Boston: PWS-Kent Publishing Co.

Salton, G. (1972). Automatic thesaurus construction for information retrieval. *Information Processing, 71,* 115–123.

Salton, G. (1978). Generation and search of clustered files. *ACM Transactions on Database Systems, 3,* 321–346.

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

Salton, G., & Lesk, M. E. (1971). Information analysis and dictionary construction. In G. Salton (Ed.), *The Smart retrieval system—Experiments in automatic document processing* (pp. 115–142). Englewood Cliffs, NJ: Prentice-Hall.

Schatz, B. (1991/1992). Building an electronic community system. *Journal of Management Information Systems, 8,* 87–107.

Schatz, B. (1993). Building collaboratories for molecular biology. In *National collaboratories: Applying information technology for scientific research.* Washington, DC: National Research Council, National Academy Press.

Smeaton, A. F., & van Rijsbergen, C. J. (1983). The effects of query expansion on a feedback document retrieval system. *Computer Journal, 26,* 239–246.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval.* London: Butterworths.

Star, S. L. (1991). *Organizational aspects of implementing a large-scale information system in a scientific community.* University of Arizona Community Systems Laboratory, Report from summer fieldwork of Worm Community System evaluation/implementation, 1991.

Wood, W. B. (1988). *The nematode Caenorhabditis Elegans.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.