A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System

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

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

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

This research presents an algorithmic approach to addressing the vocabulary problem in scientific information retrieval and information sharing, using the molecular biology domain as an example. We first present a literature review of cognitive studies related to the vocabulary problem and vocabulary-based search aids (thesauri) and then discuss techniques for building robust and domain-specific thesauri to assist in cross-domain scientific information retrieval. Using a variation of the automatic thesaurus generation techniques, which we refer to as the *concept space* approach, we recently conducted an experiment in the molecular biology domain in which we created a *C. elegans* worm thesaurus of 7,657 worm-specific terms and a *Drosophila* fly thesaurus of 15,626 terms. About 30% of these terms overlapped, which created vocabulary paths from one subject domain to the other. Based on a cognitive study of term association involving four biologists, we found that a large percentage (59.6–85.6%) of the terms suggested by the subjects were identified in the conjoined fly-worm thesaurus. However, we found only a small percentage (8.4–18.1%) of the associations suggested by the subjects in the thesaurus. In a follow-up document retrieval study involving eight fly biologists, an actual worm database (Worm Community System), and the conjoined fly-worm thesaurus, subjects were able to find more relevant documents (an increase from about 9 documents to 20) and to improve the document recall level (from 32.41 to 65.28%) when using the thesaurus, although the precision level did not improve significantly. Implications of adopting the concept space approach for addressing the vocabulary problem in Internet and digital libraries applications are also discussed.

1. Introduction

The *vocabulary (difference) problem* in human-computer interactions has been studied extensively in recent years (Furnas, 1982; Furnas, Landauer, Gomez, & Dumais, 1987). Furnas et al. (1987) found that in spontaneous word choice for objects in five domains, two people favored the same term with less than 20% probability. This fundamental property of language limits the success of various design methodologies for vocabulary-driven interaction. In information science, indexing and search uncertainty have been recognized as the primary sources of information retrieval problems. Previous research (Bates, 1986) has shown 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 (possibly because of learning or the cognitive state of mind at indexing). A high degree of uncertainty with regard to search terms has also been reported: Searchers tend to use different terms for the same information sought. Because of the indeterminism involved in indexing and searching, an exact match between the searcher's terms and those of the indexer is unlikely (Chen & Dhar, 1987). This often results in poor recall and precision in search.

The recent Human Genome Initiative (HGI) offers tremendous challenges not only to the biology, biomedi-

Received March 1, 1995; revised August 2, 1995; accepted November 2, 1995.

© 1997 John Wiley & Sons, Inc.

cine, and genetics research communities, but also to the information science and computer science communities. According to Courteau (1991), the Human Genome Project "will generate more data than any single project to date in biology," resulting in complete sequences and physical maps containing the location of every gene of the human genome and the genomes of other model organisms. Biologists study organisms in order to develop a generalizable understanding of the processes of life. The information learned about each animal is shared and compared, leading to a fuller, broader, and more detailed picture. New methods in biotechnology facilitate researchers' gathering of data at the finest levels of granularity. Numerous genomes are currently being mapped and sequenced, including those of nematode worm, fruit fly, mouse, man, bacteria (*E. coli*), yeast, loblolly pine, triticale wheat, and others. Because research communities in biology tend to form around organisms, rather than phenomena or processes, separations between communities generally indicate not only distinct groups of people, but distinct databases and vocabularies.

The vocabulary problem caused by the nomenclature and semantic differences between biological subdomains complicates the problem of information retrieval and sharing. While there is common terminology among the various subdisciplines for biological concepts (e.g., cellular functions), names for genes, physiological functions, and anatomical parts can differ from species to species. Nomenclature schemes and naming conventions vary widely among the different biological research communities. Some, such as those of the very young worm community, are highly standardized. In contrast, others, including the yeast and fly domains, involve very little standardization. Terms can also have different semantic meanings in various biological systems. For example, in the nematode sperm are pseudo pods that crawl; in other systems, these are ciliated flagella that swim. In addition, the language of science is highly dynamic and fluid over time (Frenkel, 1991). Not only does the vocabulary change to represent increased understanding as scientists continue to learn about the systems they study, but old terms can take on broader, narrower, or even different meanings as research advances.

The vocabulary problem in scientific research demands the development of advanced computing techniques. One recent attempt to address the problem vocabulary differences in molecular biology research is the development of the Worm Community System (WCS) as part of the NSF *National Collaboratory* effort (Rosenberg, 1992; Schatz, 1993). This experiment in building an electronic scientific community system for the *C. elegans* biologists has been considered a model electronic community system (Pool, 1993). It offers traditional database functionalities along with literature, informal information and research lore, mapping programs and graphics, and allows 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, 1991/1992). The current WCS runs under X-Windows on Unix machines and can also be used remotely from Sun and DEC workstations and Macintosh personal computers (Shoman, Grossman, Powell, Jamison, & Shatz, 1995).

In order to address the vocabulary problem in information retrieval for worm biologists (both experts and novices), we developed and integrated into the WCS an automatically generated thesaurus containing domain-specific vocabulary related to the worm (Chen, Schatz, Yim, & Fye, 1995). In response to a searcher's query, the thesaurus component suggests related worm concepts that serve to trigger the searcher's recognition and thereby broaden or sharpen the search. The present work involves development and evaluation of a second automatic thesaurus for the domain of *Drosophila melanogaster* (fruit fly) genetics and molecular biology, with the goal of integrating this with the worm thesaurus. We believe that the conjoined fly-worm thesaurus and their overlapping vocabularies could suggest meaningful *vocabulary paths* to lead community outsiders (e.g., fly biologists) into a different subject domain and identify research documents (e.g., worm literature) of interest. This research proposed an algorithmic and scalable approach to cross-domain scientific information retrieval, using WCS and the worm-fly biology community as an experimental testbed.

In Section 2, we present a review of vocabulary association studies and thesaurus work. A *concept space* approach, which is grounded on cluster analysis and general AI search algorithms, is then presented in Section 3. We also summarize our previous findings in the same section. In Section 4, we present a cognitive study which investigated the concept (term) association behaviors of four biologists who are knowledgeable about both fly and worm genetics. Section 5 presents a follow-up study which involved eight fly biologists who were asked to retrieve worm documents in the WCS, with and without the help of the fly-worm thesaurus. Both experiments included quantitative measures, statistical analysis, and (verbal) protocol analysis. Conclusions and a discussion are presented in the last section.

2. Vocabulary Association and Thesaurus: Literature Review

The vocabulary problem affects every domain of human knowledge. Based on research over the past few decades, it has become clear to information scientists that development of effective online information retrieval systems must consider the cognitive processes and the vocabulary association characteristics of the users (Chen & Dhar, 1991).

2.1. *Vocabulary Association*

According to Belkin, users of information retrieval systems bring with them a problem statement which represents an information need. Inherent in all information needs are "anomalous states of knowledge" (ASKs) (Belkin, Oddy, & Brooks, 1982a). In Belkin's document retrieval system based on ASKs (Belkin et al., 1982a, 1982b), the searcher's state of knowledge is represented as a network of associations between words. From the structure and characteristics of the network, it is possible to identify anomalies in the state of knowledge. The ASKs model has also contributed to associative indexing and term-association based retrieval. Belkin's research shows that "networks constructed from constrained word associations yield reasonable representations of individuals' states of knowledge about the subject to which the associations are constrained." The cognitive experiment reported in this article was also based on such a word-association network representation.

Several models of human memory association have been suggested wherein knowledge is represented by network-like structures with linked propositions. Anderson's work in human memory is particularly pertinent to term associations in retrieval (Anderson, 1985a, 1985b). According to Anderson, people remember not the exact wording of verbal communication, but the meaning underlying it. The smallest unit of knowledge that can stand as an assertion bearing meaning is the proposition. Memory, then, is represented as a network of such propositions. The strength of the association paths leading to a particular piece of information contributes to the level of activation being spread. This theory of *spreading activation* has influenced the design of many semantic network-based information retrieval systems (Chen & Ng, 1995; Cohen & Kjeldsen, 1987; Shoval, 1985).

2.2. *Thesaurus Work*

Information retrieval in large document collections often requires vocabulary expansion aids because, as Blair and Maron (1985) contend, "vocabulary problems make high recall impossible in full-text databases." Furnas et al. (1987) and Gomez, Lochbaum & Landauer (1990) found in their studies that "searcher success is markedly improved by greatly increasing the number of names per object." They proposed an "unlimited aliasing" strategy, which allows essentially unlimited numbers of aliases for objects, to alleviate the vocabulary difference problem.

Many research groups have created vocabulary-based search aids for online information retrieval systems by making use of existing thesauri or dictionaries. Thesauri, in particular, exhibit a structure similar to human word-association networks. While these tools are able to provide the searcher with 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 thesauri or dictionaries in the first place. An alternative approach to creating vocabulary-based search aids is based on *automatic thesaurus generation.*

• *Incorporating Existing Thesauri.* The National Library of Medicine's *Unified Medical Language System (UMLS)* project is one of the largest-scale efforts adopting existing domain-specific knowledge sources or thesauri in information access. It 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 (consisting of biomedical concepts and their relationships as presented in more than 10 different existing vocabularies and thesauri); a Semantic Network (continuing information about and relationships between the categories or classes included in the Metathesaurus); and an Information Sources Map or directory (containing information about various biomedical databases). The system suggests terms for user selection.

Many recent information science projects also have adopted multiple existing thesauri for term suggestion. Chamis discussed the issues of thesaurus compatibility and strategies and systems developed to overcome difficulties in searching multiple incompatible databases (Chamis, 1991). In particular, she 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 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 of these various microthesauri, 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 developed an integrated vocabulary for the energy domain which represented terms from 11 existing vocabularies (Niehoff, 1976; Niehoff & Kwansy, 1979).

Several projects have attempted to incorporate existing thesauri in the design of knowledge-based information retrieval systems. Fox et al. focused on creation of so-called "relational thesauri." For example, Fox's CODER system adopted the *Handbook of Artificial Intelligence* and *Collin's Dictionary* (Fox, 1987; Fox, Nutter, Ahlswede, Evens, & Markowitz, 1988). Ahlswede and Evens (1988) parsed *Webster's Seventh New Collegiate Dictionary* to obtain a "lexical database" containing lexical or lexical-semantic relationships from the dictionary definitions. Lesk converted an online version of Murray's *Oxford English Dictionary* into a thesaurus-like tool to facilitate searching of historical manuscripts. These approaches represent attempts to produce "universal lexicons," rather than domain-specific thesauri or dictionaries. Chen et al. conducted a series of experiments which included several large-scale, domain-specific thesauri. Chen and Dhar (1991) incorporated a portion of the LCSH in the computing area into a system that used a branch-and-bound spreading activation algorithm to assist users in query formulation. More recently, they developed concept-based document retrieval using multiple thesauri: Two existing thesauri (LCSH and the ACM Computing Review Classification System) and an automatically-generated computing-specific thesaurus (Chen, Lynch, Basu, & Ng, 1993; Chen & Ng, 1995).

• *Automatic Thesaurus Generation.* Numerous investigators have developed algorithmic approaches to *automatic thesaurus generation.* Most of these approaches employ techniques that compute coefficients of "relatedness" between terms 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).

Stiles (1961) was one of the early researchers to report improved retrieval performance using a method based on term association (with collections of librarian-applied subject tags). Doyle (1962) further argued that the principles underlying association-based retrieval should apply whether the associations are determined by humans or by machines (programs). Courtial and Pomian (1987) argued that searches performed in the realms of science and technology frequently involve association of concepts that lie outside the traditional associations represented in thesauri. Associative networks gleaned through textual analysis, they argued, facilitated innovation by making obvious associations that would otherwise be impossible for humans to find on their own. In early research, Lesk (1969) found little overlap between term relationships generated through term associations and those presented in existing thesauri.

More recently, Crouch and Yang (1992) automatically generated thesaurus classes from text keywords, 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). Ekmekcioglu, Robertson, & Willet (1992) tested retrieval performances for 110 queries on a database of 26,280 bibliographic records using four approaches: Original queries and query expansion using co-occurrence data, Soundex code (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 were no significant differences 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 if a searcher can select and use the terms suggested by a co-occurrence thesaurus in addition to the terms he/she has generated.

The limitation of the popular symmetric similarity functions, e.g., cosine, Dice, and Jaccard's, have been reported by Peat and Willett (1991). Their research showed that similar terms identified by symmetric co-occurrence function tended to occur very frequently in the database that is 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' 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.

Several research groups have experimented with an algorithmic approach to cross-domain term switching recently. Chen et al. experimented extensively in generating, integrating, and activating multiple thesauri (some were existing thesauri, others automatically generated, all in computing-related areas) (Chen et al., 1993; Chen & Ng, 1995). Both Kim and Kim (1990) and Chen et al. (1993) proposed treating (automatic and manually-created) thesauri as a neural network or semantic network and applying spreading activation algorithms for term-switching. Despite questions about the usefulness of automatic thesaurus browsing heuristics (Jones et al., 1995), our recent experiment revealed that activation-based term suggestion was comparable to the

manual thesaurus browsing process in document recall and precision, but that the manual browsing process was much more laborious and cognitively demanding (Chen & Ng, 1995).

3. The Concept Space Approach to Automatic Thesaurus Generation: Techniques and Results

After closely examining past research (both in information science and cognitive studies) and based on our own experience in creating domain-specific thesauri in several scientific, engineering, and business domains, we believe that creating robust and useful domain-specific thesauri (not universal thesauri) automatically is feasible and these thesauri can potentially pave the way for cross-domain scientific information retrieval. We refer to our approach to automatic thesaurus generation as a *concept space* approach because our goal is to create a meaningful and understandable *concept space* (a network of terms and weighted associations) which could represent the concepts (terms) and their associations for the underlying *information space* (i.e., documents in the database).

3.1. Techniques: The Concept Space Approach

The specific steps and algorithms adopted include: *Document and object list collection, object filtering and automatic indexing, co-occurrence analysis,* and *associative retrieval.* We present below a brief overview of these techniques in the context of our fly-worm experiment. For algorithmic details, readers are referred to (Chen, Hsu, Orwig, Hoopes, & Nunamaker, 1994a; Chen & Lynch, 1992; Chen et al., 1993; Chen & Ng, 1995; Chen, Schatz, Martinez, & Ng, 1994b; Chen et al., 1995).

• *Document and Object List Collection.* In any automatic thesaurus building effort, the first task is to identify complete and recent collections of documents in specific subject domains that can serve as the sources of vocabularies. The proliferation of Internet services and the availability of online bibliographic databases have made document collection much easier.

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. 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' directories), and other domain-specific objects (e.g., genes, experimental methods, organizational names, etc.) which exist online or can be obtained through OCR scanning. These domain-specific keywords can be used to help identify important concepts in documents automatically.

In creating a worm thesaurus, we collected documents from four sources: The Worm Book (a reference book used widely by worm biologists, with 12 review chapters and about 700 pages of text), journal abstracts (1,467 articles, acquired from Medline and Biosis), Worm Breeder's Gazette (worm newsletter, 1,626 documents dating back to 1974), and conference proceedings articles (1,313 documents, 1977–1992). Lists of researcher names, gene names, experimental methods, and subject descriptors were also created from existing online sources. For this young and limited molecular and genetics domain, our collections (identified with the help of several worm biologists at the Arizona Worm Lab) were considered complete. On the other hand, the *Drosophila* community is one of the oldest groups in biological research. We were able to collect only recent online documents for thesaurus generation: 5,854 abstracts from Medline and Biosis (1983–1993). However, we were able to obtain four large online lists: Gene names, function names, researcher names, and subject descriptors from *FlyBase* (a set of linked databases about fly research, maintained by the Department of Biology at Indiana University). These vocabulary sources were also identified with the help of various fly biologists.

• *Object Filtering and Automatic Indexing.* For each online document, we first identified terms that matched with terms in our known vocabularies, a process referred to as *object filtering.* Because after object filtering the remaining texts may still contain many important concepts, an automatic indexing procedure then was followed. 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 non-semantic 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.

Since our first worm experiment (Chen et al., 1995), we have made several changes in the above automatic indexing process and have fine-tuned our algorithms according to subjects' suggestions. We removed the stemming procedure from automatic indexing in order to avoid creating noise and ungrammatical phrases, e.g., cloning will not be stemmed as clone (one is a process, the other is an output), *C. elegans* will not be stemmed to *C. elegan,* which is ungrammatical, etc. We created a separate 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 alleles for genes since a gene and a gene with allele have different meanings, e.g., daf-9 and daf-9(e1406). We believe these revisions were essential for identifying specific biological concepts and creating precise and useful thesauri.

• *Co-occurrence Analysis.* After terms were identified in each document, 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. A few changes were made to the standard *term frequency* and *inverse document frequency* measures.

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 object 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 object filtering were assigned heavier weights than terms identified by automatic indexing.

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 convey more precise semantic meaning than single-word terms.

We then performed term co-occurrence analysis based on the asymmetric "Cluster Function" developed by Chen and Lynch (1992). We have shown that this asymmetric similarity function represents term association better than the popular cosine function. The weighting-factor appearing in the equations below is a further improvement of our cluster algorithm.

$$ClusterWeight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times Weighting\ Factor(T_k)$$

$$ClusterWeight(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times Weighting\ Factor(T_j)$$

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 descriptors 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.

In order to *penalize* general terms (terms which appeared in many places) in the co-occurrence analysis, we developed the following weighting schemes which are similar to the *inverse document frequency* function:

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

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

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, e.g., *C. elegans*, development, 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. After consulting worm and fly biologists in the Arizona worm and fly laboratories 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.

• *Associative Retrieval.* In addition to the user-controlled thesaurus browsing process, searchers can also invoke selected spreading activation algorithms for multiple-term, multiple-link term suggestions. We have developed two algorithms, based on the serial branch-and-bound algorithm and the parallel Hopfield net algorithm, respectively (Chen & Ng, 1995). The Hopfield algorithm, in particular, has been shown to be ideal for concept-based information retrieval.

The Hopfield (1982) net was introduced as a neural network that can be used as a content-addressable memory. Knowledge and information can be stored in single-layered, inter-connected neurons (nodes) and weighted synapses (links) and can be retrieved based on the Hopfield network's *parallel relaxation* and *convergence* methods. The Hopfield net has been used successfully in such applications as image classification, character recognition, and robotics (Knight, 1990; Tank & Hopfield, 1987) and was first adopted for *concept-based* information retrieval in Chen et al. (1993).

Each term in the network-like thesaurus was treated as a neuron and the asymmetric weight between any two terms was taken as the unidirectional, weighted connection between neurons. Using user-supplied terms as input patterns, the Hopfield algorithm activated their neighbors (i.e., strongly associated terms), combined weights from all associated neighbors (by adding collective association strengths), and repeated this process until convergence. During the process, the algorithm caused a *damping effect*, in which terms farther away from the initial terms received gradually decreasing activation weights and activation eventually "died out." This phenomenon is consistent with the human memory *spreading activation* process.

The Hopfield net algorithm relied on an activate and iterative process, where

$$\mu_j(t+1) = f_s\left[\sum_{i=0}^{n-1} t_{ij}\mu_i(t)\right], \quad 0 \le j \le n-1$$

$\mu_j(t+1)$ is the activation value of neuron (term) j at iteration $t+1$, t_{ij} is the co-occurrence weight from neuron i to neuron j, and f_s is the continuous SIGMOID transformation function, which normalizes any given value to a value between 0 and 1 (Dalton & Deshmane, 1991; Knight, 1990). This formula shows the *parallel relaxation* property of the Hopfield net. [Readers are referred to (Chen & Ng, 1995) for algorithmic detail.]

The experiments reported in this research did not contain the *associative retrieval* component. As a first step toward verifying the concept space approach to alleviating the vocabulary problem in scientific retrieval, we provided only a graphical user interface for browsing the fly and worm thesauri created for the Worm Community System. Our ongoing work involves incorporating the associative retrieval component in several large-scale, operational systems (Chen & Schatz, 1994).

3.2. Prior Results: Worm and Fly Thesauri

By adopting the *concept space* approach and working closely with worm and fly biologists in the Molecular and Cellular Biology (MCB) Department at the University of Arizona for about 2 years, we generated a worm thesaurus in Fall 1993 (Chen et al., 1995) and a fly thesaurus in Summer 1994 (Chen et al., 1994b). Both thesauri had been independently tested by the biologists and are available for Internet WWW access (*BioQuest*) at: http://ai.bpa.arizona.edu/.

The resulting worm thesaurus consisted of 7,657 terms and 547,810 links and the fly thesaurus contained 15,626 terms and 750,314 links (after applying various thresholds). Most of these terms were author names or subject descriptors. It took 50 and 70 minutes, respectively, to generate the two thesauri on a DEC Alpha 2100 workstation (200 MHz, 128-MB RAM). The resulting thesauri were about the same size as the initial document collections (i.e., 1:1 storage overhead).

A structural analysis of the two thesauri revealed that about 30% of their subject descriptors overlapped (Table 1). Not surprisingly, we found little overlap in author or gene names. Overall, about 10% of the fly terms overlapped with worm terms and about 21% of the worm terms overlapped with fly terms. These overlapping terms provided potentially useful "vocabulary paths" from one domain to the other.

Based on the two automatic thesauri created for worm and fly biology, we proceeded to test their usefulness for cross-domain concept-based retrieval. The first experiment aimed to understand fly-worm biologists' (biologists who are familiar with both worm and fly biology) cross-domain term association patterns and their similarity to the terms and associations represented by the fly-worm thesaurus. The second experiment involved implementing the thesauri (and a graphic user interface) on the operational Worm Community System and investigating the retrieval performances (recall, precision, and subjective evaluation) of fly biologists when using the conjoined thesaurus to help retrieve worm documents (i.e., using fly terms to retrieve worm documents).

4. Fly-Worm Thesaurus Traversal Experiment

4.1. Experimental Design

The goal of this experiment was to understand fly and worm biologists' associations between concepts—associations that form the basis for the decisions and inferences they use when searching information. Four subjects from the fly and worm domains were asked to identify paths of associated terms that might be taken to traverse from terms in one domain to terms in the other domain. The fly subjects were both faculty members; the worm subjects were both graduate students. Subjects identified pairs of terms—one term from the fly domain, one from the worm domain—that they knew to have equivalent semantic meaning in the two domains. They were asked to articulate clearly any thoughts that occurred to them as they developed their network of associations. While discussing term associations and introducing new associated terms that link the two initial terms, subjects drew graphs depicting concept relationships. Verbal protocols

TABLE 1. Number of overlapping terms between fly and worm thesauri.

	Fly			Worm	
Object	Total fly terms	% Overlapping with worm	Overlap common terms	% Overlapping with fly	Total worm terms
Author	8,153	3.21	262	12.52	2,092
Function	224	14.29	32	100.00	32
Gene	3,315	0.39	13	1.54	845
Subject	3,935	30.93	1,267	27.03	4,688
Total	15,626	10.08	1,574	20.56	7,657

were tape recorded and transcribed for subsequent analysis.

Terms (nodes) and associations (links) expressed by the subjects were searched in the conjoined fly-worm thesaurus to determine how many appeared. Counting was done for both partial (subset) and whole phrases. Also, the networks drawn by the subjects were analyzed to elucidate traversal behavior and strategies. The four subjects completed a total of 18 traversal graphs between fly and worm terms. The time required for the experiment ranged from 35 minutes for one expert who completed five traversals, to 1 hour 30 minutes for one graduate student who completed three traversals.

4.2. A Sample Traversal and Analysis of Traversal Graphs

Figure 1 illustrates the process taken in analyzing the traversal graphs. Panel (a) depicts the graph of terms as it was drawn by the subject. The time (in minutes and seconds after the beginning of the traversal) at which each term was stated by the subject is noted beneath each term. The order of traversal may be determined by following the passage of time. The source node was defined as the term in the initial term pair that came from the subject's domain and the target node was the term in the other domain. Intermediate nodes were terms that were used in traversing between the two domains.

The subject first identified the source term (let-23) and target term (sevenless) (both gene names) and then proceeded to list commonalities between the two genes. They are both *cell fate determinants* within *signal transduction pathways* and they encode *receptor protein kinases*, which are located on the *cell surface* [see the terms in the middle of panel (a)]. The subject then named closely associated proteins, *boss* and *ras* (in the two domains: *Dras* and *let-60*, respectively). Next, the subject stated that the genes function in the development process of different tissues in the two animals: *Vulval development* in *C. elegans*, and *eye development* in *Drosophila* (specifically in the *R7 cell determination*). Finally, the subject summarized the relationship between the two genes: They perform similar functions using similar *mechanisms*, but do so in different *tissues*.

In panel (b), all terms not directly involved in a traversal to terms in the fly domain have been removed. In this case, only the associated gene *boss* was removed from the graph. Also, the links to the two *ras* nodes have been altered to create a path that extends from let-23 to sevenless. The number of terms removed from each graph analyzed depended upon the extent to which subjects discussed domain-specific details about the initial terms. The resulting graph depicts a variety of paths that could be taken to traverse from one domain to the other. We then searched the conjoined fly-worm thesaurus for all terms and associations included in panel (b).

Panel (c) depicts the nodes and links found in the fly-worm concept space. Nodes found have been marked according to the object type: Gene name (oval) and subject term (rectangle). Terms existing in both the fly and the worm concept space are enclosed in a large circle. Terms to the left of the large circle were found only in the worm domain and those to the right were found only in the fly domain. All but one gene (Dras) were found in the thesaurus. While the whole phrase for subject terms may not have been found in the concept space, component words of suggested term phrases were found. For example, in Figure 1, components of all term phrases were found, but only "vulval development" was found as a complete term phrase.

4.3. Experimental Results: Matching Terms and Associations in Thesaurus

The majority of suggested terms were subjects, which were mostly multiple word phrases. Biologists have several ways of referring to the same concept, depending upon the level of specificity they wish to convey in a given discussion. One example of this would be the phrase "receptor tyrosine kinase." Other acceptable names for the same concept would be "receptor kinase" or "tyrosine kinase." All are essentially synonymous. Due to variations in statements of concepts, it was necessary to compute the statistics for the number of suggested terms that exist in the thesaurus in two ways: By searching for the whole phrase as suggested by the subject and by searching for the various component words and phrases making up the suggested phrase. Results of our analysis are summarized below:

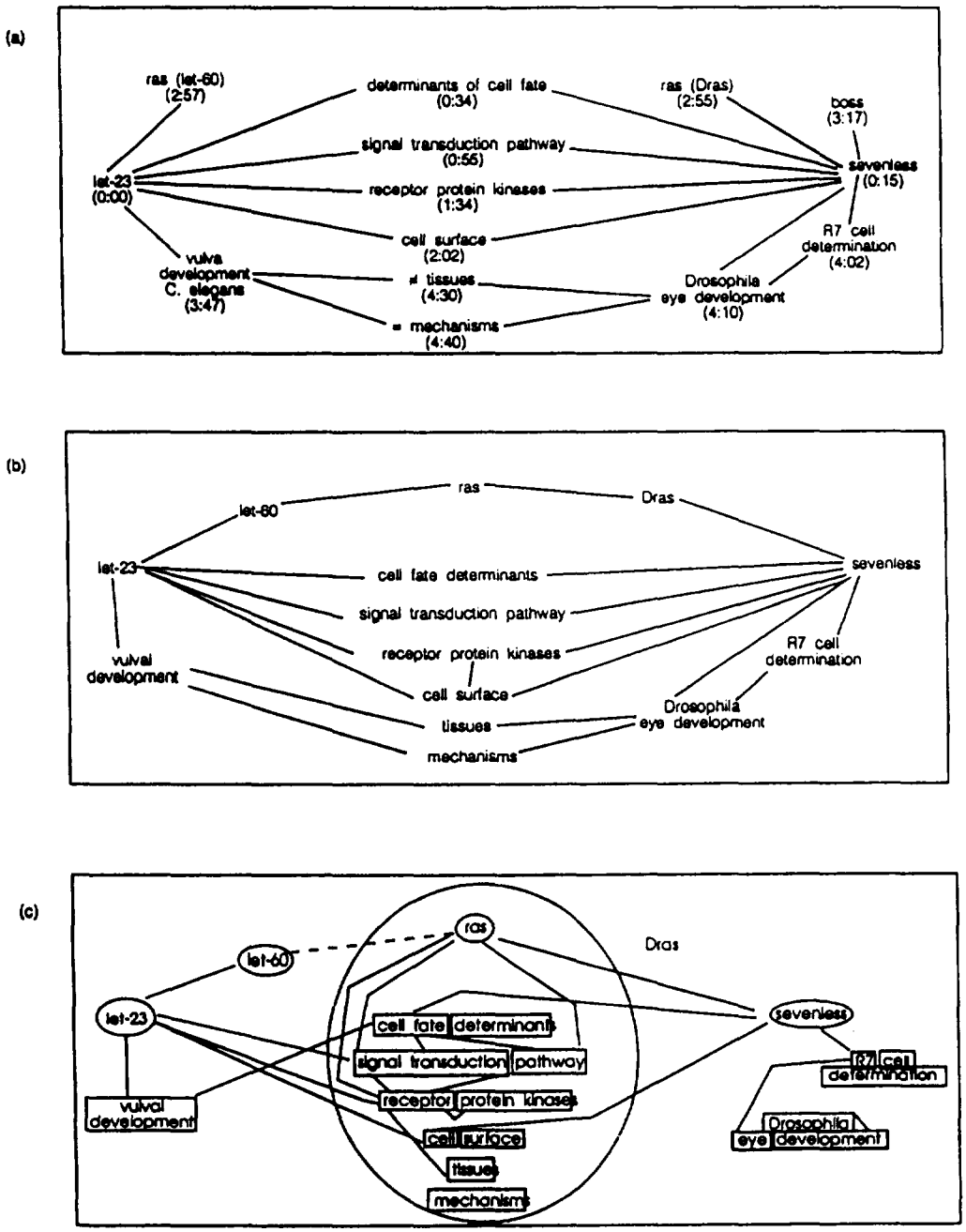

FIG. 1. a–c: Let-23—sevenless traversal.

• *A Large Percentage of the Worm and Fly Terms Were Found in the Thesaurus.* Tables 2 and 3 show the number of biologist-suggested terms identified in the conjoined fly-worm thesaurus (i.e., Found/Suggested). A greater percentage of worm-specific than fly-specific terms were found in the respective thesauri, regardless of either the domain-affiliation of the subject or the manner of phrase searching (whole phrase vs. partial phrase). This is likely to have resulted from the difference in completeness of the two thesauri. The worm collection was significantly more complete than the fly collection, as discussed earlier.

Overall, 59.6% of the (whole) phrases suggested by the subjects were found in the thesaurus. In contrast, 85.6% of the component (partial) phrases were identified. For a total of 146 terms stated by the biologists, subject descriptors and gene names comprised the majority of the concepts.

• *A Small Percentage of the Term Associations Were Found in the Thesaurus.* Table 4 shows the result of searching the conjoined fly-worm thesaurus for links. Associations suggested by subjects were again counted for both whole phrases and partial phrases (subsets). Based on whole phrases suggested by subjects, 8.4% of links were found in the conjoined thesaurus (for a total

TABLE 2. Number of nodes (whole phrases) found in conjoined fly-worm thesaurus.

Subject	Object	Worm specific found/suggested	Fly specific found/suggested	Common term found/suggested	Total found/suggested	% Found
Worm experts	Author	2/2	1/1	0/0	3/3	100.0
	Function	0/0	0/0	0/0	0/0	
	Gene	8/8	6/11	0/0	14/19	73.7
	Subject	1/4	1/3	26/47	28/54	51.8
	Total	11/14	8/15	26/47	45/76	59.2
	Percent	78.6%	53.3%	55.3%	59.2%	
Fly experts	Author	0/0	0/0	0/0	0/0	
	Function	0/0	0/0	0/0	0/0	
	Gene	8/8	8/10	2/9	18/27	66.7
	Subject	6/6	2/6	16/31	24/43	55.8
	Total	14/14	10/16	18/40	42/70	60.0
	Percent	100.0%	62.5%	45.0%	60.0%	
Overall	Author	2/2	1/1	0/0	3/3	100.0
	Function	0/0	0/0	0/0	0/0	
	Gene	16/16	14/21	2/9	32/46	69.6
	Subject	7/10	3/9	42/78	52/97	53.6
	Total	25/28	18/31	44/87	47/146	59.6
	Percent	89.3%	58.1%	50.6%	59.6%	

of 381 links/associations). When searching the thesaurus using partial phrases, 18.1% of possible links to other subsets were identified (for a total of 543 links/associations). This indicated a difference between the way terms were associated in the biologists' long term memory and the way they were associated in the thesaurus. However, the thesaurus may have served as an additional query expansion aid to augment a biologist's long-term memory. (This hypothesis was tested in the next experiment.)

• *Terms Associations Were Bidirectional.* Finally, we considered the directionality of links, comparing links flowing from domain-specific terms to common terms and vice versa. We found about equal proportions of associations from common terms to domain-specific terms and from domain-specific terms to the common terms. This indicated a bidirectional nature of term associations for cross-domain concepts.

4.4. Experimental Results: Traversal Behaviors

The traversal graphs and verbal protocols were analyzed to determine subjects' heuristics for traversing from one domain into the other.

TABLE 3. Number of nodes (partial phrases) found in conjoined fly-worm thesaurus.

Subject	Object	Worm specific found/suggested	Fly specific found/suggested	Common term found/suggested	Total found/suggested	% Found
Worm experts	Author	2/2	1/1	0/0	3/3	100.0
	Function	0/0	0/0	2/2	2/2	100.0
	Gene	8/8	6/11	0/0	14/19	73.7
	Subject	6/8	5/5	61/68	72/81	88.9
	Total	11/14	8/15	26/47	91/105	86.7
	Percent	88.9%	73.5%	90.0%	86.7%	
Fly experts	Author	0/0	0/0	0/0	0/0	
	Function	0/0	0/0	0/0	0/0	
	Gene	8/8	8/10	2/9	18/27	66.7
	Subject	6/6	10/10	61/68	77/84	91.7
	Total	14/14	18/20	61/68	95/111	85.6
	Percent	100.0%	90.0%	89.7%	85.6%	
Overall	Author	2/2	1/1	0/0	3/3	100.0
	Function	0/0	0/0	2/2	2/2	100.0
	Gene	16/16	14/21	2/9	31/46	69.6
	Subject	12/14	15/15	122/136	149/165	90.3
	Total	30/32	30/37	124/147	185/216	85.6
	Percent	93.8%	81.1%	84.3%	85.6%	

TABLE 4. Number of suggested links found in conjoined thesaurus.

Subject	Link	Partial phrases		Whole phrases	
		Subject suggested	Additional found	Subject suggested	Additional found
Worm experts	Fly-fly	2/7	0	0/7	3
	Fly-common	6/56	2	2/33	0
	Common-fly	9/63	2	2/36	3
	Common-common	0/0	60	3/39	3
	Common-worm	13/57	1	3/36	5
	Worm-worm	5/10	1	4/7	4
	Worm-common	12/66	0	2/36	3
	Total	47/259	66	15/182	21
	Percent	18.1%		8.2%	
Fly experts	Fly-fly	3/10	1	2/6	0
	Fly-common	15/67	4	3/38	0
	Common-fly	7/72	6	4/42	1
	Common-common	0/0	43	1/33	8
	Common-worm	14/63	4	1/39	1
	Worm-worm	2/8	4	4/9	1
	Worm-common	12/64	4	2/32	1
	Total	53/284	66	17/199	21
	Percent	18.7%		8.5%	
Overall	Fly-fly	5/17	1	2/13	3
	Fly-common	21/123	6	5/71	0
	Common-fly	16/135	8	6/78	4
	Common-common	0/0	103	4/72	6
	Common-worm	27/120	5	4/75	6
	Worm-worm	7/18	5	8/16	5
	Worm-common	24/130	4	4/68	4
	Total	100/543	133	32/381	33
	Percent	18.1%		8.4%	

• *Most Traversals Used Only One Intermediate Node.* Both fly and worm experts generally used just one intermediate node when traversing between the two domains: 66% of worm subjects' traversals and 72% of fly subjects' traversal. Overall, 69% of traversals used one intermediate term, 13% used two intermediate terms, and 18% used 3–5 intermediate terms. The worm subjects performed a greater number of searches using two or three intermediate nodes than did fly subjects (31% compared to 14%). It appeared that the biologists' term *spreading activation* often involved limited levels of links, i.e., 2–3 links for the majority of the cases.

• *Terms Associations Were Context-Driven.* In creating associations between related terms, subjects often pointed out specific similarities and/or differences between the two initially identified (source and target) terms. Based on our protocol analysis, we found that several contexts for these similarities and differences existed, including, two genes were identified by similar (or different) experimental strategies; their cellular structures had similar (or different) composition; two proteins were involved in similar or different cellular or developmental processes; genes manifested similar or different phenotypes; genes or proteins had similar or dissimilar sequences (homology) or contained similar motifs or domains; proteins or genes performed similar (or dissimilar) functions; two genes were members of the same gene family or involved in the same type of pathway; and two genes existed or functioned in the same or different cell types.

• *Stories of Historical Development Were Important for Association.* Biologists looked to other domains for hints as to what might be happening in their own domain. Several protocols included "stories" of historical development of the current understanding about genes, proteins, processes, etc. The importance of timely information exchange in the advancement of biology was exemplified by one of the experts who, in distinguishing between the two phenomena he was discussing, indicated that the particular function had been "shown" to be true in one domain, but was only "hypothesized" to be true in the other domain.

In summary, we felt that the results of the term association experiment were very encouraging. The high probability of occurrences of subject-supplied terms in the conjoined fly-worm thesaurus indicated a strong likelihood that users can "dock" onto to the concept space easily [using Bates's (1986) terminology]. However, the association links suggested by the thesaurus were often different from those provided by the subjects. The use-

fulness of the thesaurus associations needed to be investigated, especially for cross-domain scientific information retrieval.

5. Fly-Worm-WCS Document Retrieval Experiment

5.1. Experimental Design

With the encouraging results obtained from the traversal experiment, we proceeded to integrate the conjoined fly-worm thesaurus into the Worm Community System and conducted a follow-up document retrieval experiment. The goal of this experiment was to find out whether a conjoined thesaurus, representing conceptual associations in two related but distinct subdomains of the biological research community, was able to bridge the vocabulary differences between those subdomains and assist in cross-domain information retrieval.

Eight subjects with expertise of varying levels in the domain of *Drosophila* research defined and performed their own searches using the Worm Community System, with the aim of identifying useful or relevant worm documents in response to fly-related queries. Due to a copyright issue and the database content, WCS only contained worm documents, which permitted cross-domain search from fly queries to worm documents, but not the reverse. For comparison, each query was performed twice: First without the assistance of the conjoined thesaurus and then with the thesaurus. Subjects were encouraged to make use of the full range of keyword and Boolean searching and hypertext browsing capabilities available in the WCS. Subjects were asked to evaluate the relevance of each WCS item and document retrieved. The search session and relevant worm documents identified were recorded by an experimenter. Subjects were asked to think aloud and their verbal protocols were recorded and transcribed for subsequent protocol analysis. For determination of recall, the complete search session and output of each query were subsequently evaluated by a "super-expert" to identify a target set of relevant documents. Each subject spent between 1 and 2 hours making their queries. The super-expert, a *Drosophila* researcher (faculty) with over 10 years of experience in the field, spent almost 10 hours in reviewing all the results.

5.2. Experimental Results: Relevant Documents, Recall, and Precision

The eight subjects attempted a total of 36 queries. Twenty-two of these queries were not included in the subsequent analysis either because the WCS did not contain any document relevant to the queries or because queries were not carried through to completion by the subjects. Relevant documents retrieved and the recall and precision measures were calculated using the remaining 14 completed queries.

Thesaurus use contributed to development of more complex, and potentially more specific queries. Of the 14 queries considered in calculation of recall and precision, eleven resulted in formulation of more complex, Boolean queries. In contrast, of the 22 queries that were not considered in calculations, only four resulted in Boolean reformulations. Results of this experiment are summarized below:

- *The Conjoined Thesaurus Helped Find More Relevant Documents.* Results from calculation of relevant documents retrieved were based on the target set of relevant documents identified by the super-expert. Without the aid of the fly-worm thesaurus, searchers were able to find 8.79 relevant documents. With the assistance of the thesaurus for developing useful query terms, subjects were able to find a total of 19.93 relevant documents, almost doubling the number of documents retrieved. The additional relevant documents retrieved using the thesaurus did not often duplicate the set of documents retrieved without the use of the thesaurus. One-way analysis of variance (ANOVA) using the MINITAB statistical package (Rosenberg, 1992) showed that this improvement was statistically significant ($p = 0.059$). (In all our analysis, a 10% statistical significance level was adopted.)

- *The Conjoined Thesaurus Helped Improved Document Recall.* The number of relevant documents identified by each subject, both before and after thesaurus consultation, was determined based on the total number of relevant documents identified by our super-expert. The average recall was 32.41% without use of the thesaurus and 65.28% with the thesaurus. This improvement in recall was statistically significant ($p = 0.015$).

- *The Conjoined Thesaurus Did Not Improve Document Precision.* The overall precision was 43.51% without the thesaurus and 53.48% with the thesaurus. However, this improvement was not statistically significant ($p = 0.477$).

5.3. Experimental Results: Search Behaviors

In addition to the quantitative analysis for the experiment, verbal protocols and comments after searches were collected and analyzed. We summarize the results below:

- *Relevance Was a Subjective Concept.* Although our experiment attempted to measure retrieval performance by using standard information science measures, we often found that the concept of "relevance" is very subjective and has different meanings for different people. Even though given the same instructions, the subjects and super-expert in some cases identified different sets of documents as being relevant. Subjects identified those documents that were relevant to their information need as they understood it at the time of the search session. In contrast, the super-expert identified all docu-

ments relevant to the queries articulated by the subjects, regardless of the type of document retrieved and without knowing other unspecified constraints or visceral needs of the subjects. So even though our experimental results were positive, the retrieval behaviors of subjects in an operational environment may still vary significantly.

• *Most Queries Were about Learning Worm Biological System or Homologue.* The queries articulated by the subjects fell into several categories. However, most queries (27 out of 36) were either aimed at learning what is known about a particular biological system in the worm or at determining the name of a worm homologue for a fly gene of interest. Certain query types were more likely to result in an unsuccessful "term-switching" from fly to worm. For example, several unsuccessful search attempts were related to fly-specific functions or structures that do not exist in worms, e.g., genes or proteins related to wing function (the fly has wings, but the worm does not).

• *The Thesaurus Helped Jog Human Memory.* Many subjects were particularly impressed with the thesaurus's ability to jog their memories. Many verbal protocols supported this observation. For example, one subject said, "it triggered things in my brain. It showed me words that I knew were connected." Another expert subject reacted to a list of thesaurus terms and commented: "Oh, yeah. Definitely relevant. Definitely relevant. . . That's exactly what you would hope to be looking for." Later, in summarizing his impressions of the usefulness of the thesaurus, he referred back to that search saying, "Well, it certainly helped with the first one. I mean, you know, when we started with "wingless," and it just sort of reminded you that you should look for "wnt" as well. So, that's actually useful for that case. You still have to know enough to recognize what "wnt" is, and what it means. So it's more like a reminder than an educator in that sense. And I think that's probably one of the things that it would be used for."

Several subjects commented that a certain level of domain knowledge may be necessary in order to select appropriate terms readily. Most of the subjects were able to identify relevant terms from their own domain (fly) in the thesaurus. However several subjects, especially the junior researchers, expressed uncertainty about which worm terms offered by the thesaurus would be relevant. One subject said, "Let's try just a random gene. Let's try lin-39, and see why that came up."

• *The Thesaurus Helped Expand or Limit Queries.* Thesaurus consultation helped searchers to articulate their queries better. In most cases, subjects were better able to articulate their queries after seeing both the outcome of an initial search and the list of thesaurus-suggested terms. For example, one subject was overwhelmed when her initial query about "microtubule binding proteins" retrieved over 500 documents. After browsing through the titles, she said, "Well, those are definitely microtubule binding proteins, but they aren't the kind that I was looking for." After consulting the thesaurus, she modified the query to include two more terms. The results of the second query returned a smaller set of documents which were of interest to her.

In summary, the conjoined thesaurus had done an excellent job in helping the fly biologists find more relevant worm documents, improve search recall, jog memory, and articulate queries. However, the precision level of the searches did not improve.

6. Conclusions and Discussion

The vocabulary problem in scientific research demands the development of advanced computing techniques. This article has presented a *concept space* approach to addressing the vocabulary problem in scientific collaboration and information sharing, using the molecular biology domain as an example. We first provided a literature review of cognitive studies related to the vocabulary problem and vocabulary-based search aids. Belkin's ASKs model, which represents a searcher's state of knowledge as a network of associations between words, and Anderson's human memory model of *spreading activation* were then described to provide a theoretical foundation for query expansion in information retrieval.

Despite many positive results, numerous groups have reported poor results and even degraded performance with systems offering automatic query expansion. Based on a review of past research and our own experience in building domain-specific thesauri for various applications, we proposed a *concept space* approach to automatic thesaurus generation. The specific steps and algorithms adopted in our *concept space* approach include: *Document and object list collection, object filtering and automatic indexing, co-occurrence analysis,* and *associative retrieval.*

In an attempt to understand the usefulness and performance level of the concept space approach to addressing the information retrieval difficulties, we recently conducted an extensive experiment in the molecular biology domain. We created a *C. elegans* worm thesaurus with 7,657 worm-specific terms and a *Drosophila* fly thesaurus with 15,626 terms. About 30% of these terms overlapped, which created vocabulary paths from one subject domain to the other.

In a cognitive study of four biologists' term association, we found that a large percentage (59.6–85.6%) of the terms suggested by the subjects were identified in the conjoined fly-worm thesaurus, but that only a small percentage (8.4–18.1%) of the associations suggested by the subjects were identified in the thesaurus. Our analysis also revealed that biologists often traversed via one intermediate term and that their associations were often context-driven and story-based.

In a follow-up document retrieval study involving

eight fly biologists, the conjoined fly-worm thesaurus, and an actual worm database (Worm Community System), subjects were able to find more relevant documents (an increase from about 9 documents to 20) and document recall level improved from 32.41 to 65.28%. However, the precision level did not improve significantly. Protocol analysis also revealed that the automatic thesaurus helped jog human memory and assisted in expanding or limiting queries.

The conjoined fly-worm thesaurus has been incorporated into the Worm Community System. We also have created a scaled-down system called *BioQuest* that is available on the Internet WWW for remote access (http://ai.bpa.arizona.edu/). *BioQuest* contains several thousand documents in worm biology and allows WAIS-like keyword search and fly-worm thesaurus browsing. We are in the process of incorporating an associative retrieval component based on the Hopfield net algorithm into *BioQuest*.

As part of our ongoing NSF/ARPA/NASA funded Digital Library Initiative project, we are designing scalable algorithms for building *concept spaces* for various engineering domains (significantly larger and more complex than fly-worm biology). Several algorithms discussed earlier have been implemented on a CM-5 parallel computer (with 1024 processing units) and, recently, on the 16-node Power Challenge (both at the National Center for Supercomputing Applications at the University of Illinois). Our other ongoing work involves creating a concept space for all Internet services (homepages collected from the Lycos searchable database at the Carnegie Mellon University, http://lycos.cs.cmu.edu/), developing intelligent personal agents (spiders) based on genetic algorithms, and organizing and categorizing all Internet services using a multi-layered, graphical neural network algorithm.

7. Acknowledgments

This project was supported mainly by three grants: (1) the NSF CISE Research Initiation Award, IRI-9211418, 1992–1994 (H. Chen, "Building a Concept Space for an Electronic Community System"), (2) NSF CISE Special Initiative on Coordination Theory and Collaboration Technology, IRI-9015407, 1990–1993 (B. Schatz et al., "Building a National Collaboratory Testbed"), and (3) 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").

We would 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, we are grateful to Dr. Samuel Ward, Dr. Danny Brower, Dr. John Clark, Dr. John Little, Alicia Minniti, Lisa Werner, Bill Achazar, Dr. Lynn Manseau, John Calley, Shermali Gunawardena, Libby Heddle, Dr. Mary Rykowski, Dr. Dave Sandstrom, and Dr. Scott Selleck. We thank Ed Grossman for developing WCS and Kevin Powell for helping set up our database and computer, and keeping WCS running uneventfully during the entire experiment. We also would like to express our special thanks to Pauline Cochrane, Ann Bishop, and the referees for their valuable comments on an early draft of the paper.

References

Ahlswede, T., & Evens, M. (1988). Generating a relational lexicon from a machine-readable dictionary. *International Journal of Lexicography, 1*(3), 214–237.

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

Anderson, J. D. (1985b). Indexing systems: Extensions of the mind's organizing power. *Information and Behavior, 1*, 287–323.

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

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982a). Ask for information retrieval: Part I. Background and theory. *Journal of Documentation, 38*(2), 61–71.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982b). Ask for information retrieval: Part II. Results of a design study. *Journal of Documentation, 38*(3), 145–164.

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

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28*(3), 289–299.

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

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., & Dhar, V. (1987). Reducing indeterminism in consultation: A cognitive model of user/librarian interaction. In *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87)* (pp. 285–289), Seattle, WA, July 13–17.

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management, 27*(5), 405–432.

Chen, H., Hsu, P., Orwig, R., Hoopes, L., & Nunamaker, J. F. (1994a). 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*(5), 348–369.

Chen, H., & Schatz, B. R. (1994). Semantic retrieval for the NCSA Mosaic. In *Proceedings of the Second International World Wide Web Conference '94*, Chicago, IL, October 17–20.

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

Chen, H., Schatz, B. R., Yim, T., & Fye, D. (1995). Automatic thesau-

rus generation for an electronic community system. *Journal of the American Society for Information Science, 46*(3), 175–193.

Cohen, P. R., & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management, 23*(4), 255–268.

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

Courtial, J. P., & Pomian, J. (1987). A system based on associational logic for the interrogation of databases. *Journal of Information Science, 13,* 91–97.

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* (pp. 77–88), Copenhagen, Denmark, June 21–24.

Dalton, J., & Deshmane, A. (1991). Artificial neural networks. *IEEE Potentials, 10*(2), 33–36.

Doyle, L. B. (1962). Indexing and abstracting by association. *American Documentation, 13*(4), 378–390.

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 Educational Books.

Fox, E. A. (1987). Development of the CODER system: A testbed for artificial intelligence methods in information retrieval. *Information Processing and Management, 23*(4), 341–366.

Fox, E. A., Nutter, J. T., Ahlswede, T., Evens, M., & Markowitz, J. (1988). Building a large thesaurus for information retrieval. In B. Ballard, (Ed.), *2nd Conference on Applied Natural Language Processing, Association for Computational Linguistics* (pp. 101–108). Morristown, NJ: Bell Communications Research.

Frenkel, K. A. (1991). The human genome project and informatics. *Communications of the ACM, 34*(11), 41–51.

Furnas, G. W. (1982, March). Statistical semantics: How can a computer use what people name things to guess what things people mean when they name things. In *Proceedings of the Human Factors in Computer Systems Conference* (pp. 251–253). Gaithersburg, MD: Association for Computing Machinery.

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., & 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*(8), 547–559.

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

Hopfield, J. J. (1982). Neural network and physical systems with collective computational abilities. *Proceedings of the National Academy of Science, USA, 79*(4), 2554–2558.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J., & Walker, S. (1995). Interactive thesaurus navigation: Intelligent rules OK? *Journal of the American Society for Information Science, 46*(1), 52–59.

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.

Knight, K. (1990). Connectionist ideas and algorithms. *Communications of the ACM, 33*(11), 59–74.

Lesk, M. E. (1969). Word-word associations in document retrieval systems. *American Documentation, 20*(1), 27–38.

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* (pp. 121–125), Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November 4–7.

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* (pp. 126–130), Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November 4–7.

Niehoff, R. T. (1976). Development of an integrated energy vocabulary and the possibilities for online 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.

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*(5), 378–383.

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.

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

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

Rosenberg, L. C. (1992, January). National Science Foundation news. *SIGART BULLETIN, ACM Special Interest Group on Artificial Intelligence, 3*(1), 13–17.

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

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

Schatz, B. R. (1991/1992, Winter). Building an electronic community system. *Journal of Management Information Systems,* Special Issue.

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

Shoman, L., Grossman, E., Powell, K., Jamison, C., & Schatz, B. R. (1995). The Worm Community System, release 2.0 (WCSr2). In H. F. Epstein & D. C. Shakes, (Eds.), *Modern Biological Analysis of an Organism,* Methods in Cell Biology, Volume 48 (pp. 607–625). Orlando, FL: Academic Press.

Shoval, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing and Management, 21*(6), 475–487.

Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the Association of Computing Machinery, 8*(2), 271–279.

Tank, D. W., & Hopfield, J. J. (1987). Collective computation in neuronlike circuits. *Scientific American, 257*(6), 104–114.