A CONCEPT SPACE APPROACH TO ADDRESSING THE VOCABULARY PROBLEM IN SCIENTIFIC INFORMATION RETRIEVAL: AN EXPERIMENT ON THE WORM COMMUNITY SYSTEM

by

Joanne Martinez

A Thesis Submitted to the Faculty of the
SCHOOL OF LIBRARY SCIENCE
In Partial Fulfillment of the Requirements
For the Degree of
MASTERS OF ARTS
WITH A MAJOR IN LIBRARY SCIENCE
In the Graduate College
THE UNIVERSITY OF ARIZONA

1995

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Joanne P. Martin

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

HsinChun Chen
Asst. Professor of MIS

5/1/95
Date

Abstract

This research presents an algorithmic approach to addressing the vocabulary problem in scientific research collaboration 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 and then discuss principles and techniques for building robust and domain-specific thesauri automatically.

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.

Acknowledgments

This project was supported mainly by three NSF grants: the NSF CISE Research Initiation Award, IRI-9211418, 1992-1994 (H. Chen, "Building a Thesaurus for an Electronic Community System"), NSF CISE Special Initiative on Coordination Theory and Collaboration Technology, IRI-9015407, 1990-1993 (B. Schatz et al., "Building a National Collaboratory Testbed"), and NSF/ARPA/NASA Digital Library Initiative, 1994-1998 (B. Schatz, H. Chen, et. al, "Building the Interspace: Digital Library Infrastructure for a University Engineering Community").

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 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.

Contents

Abstract 4

Acknowledgments 5

List of Figures 8

List of Tables 9

1 INTRODUCTION 10

2 VOCABULARY ASSOCIATION AND VOCABULARY-BASED SEARCH AIDS 14

2.1 Cognitive Aspects of Vocabulary Association 14

2.2 Vocabulary-based Search Aids 17

3 THE CONCEPT SPACE APPROACH TO AUTOMATIC THESAURUS GENERATION 20

3.1 Principles: The Concept Space Approach 23

3.2 Techniques: The Concept Space Approach 32

3.3 Prior Results: Worm and Fly Thesauri 39

4 FLY-WORM THESAURUS TRAVERSAL EXPERIMENT 41

4.1 Experimental Design 41

4.2 A Sample Traversal and Analysis of Traversal Graphs 42

4.3 Experimental Results: Matching Terms and Associations in Thesaurus . 45

4.4 Experimental Results: Traversal Behavior 48

5 FLY-WORM-WCS DOCUMENT RETRIEVAL EXPERIMENT 52

5.1 Experimental Design 52
5.2 A WCS Sample Search 54
5.3 Experimental Results: Relevant Documents, Recall, and Precision 60
5.4 Experimental Results: Search Behavior 63
6 CONCLUSIONS 66
APPENDICES 69
A Experimental Instruments 69
A.1 Fly-Worm Traversal Experiment: Subject Briefing Statement 71
A.2 Fly-Worm-WCS Document Retrieval Experiment: Subject Briefing Statement 73
A.3 Fly-Worm-WCS Document Retrieval Experiment:Subject Queries Presented to Super-expert 74
B Sample Verbal Protocols 77
C Tables and Graphs 100
D Inclusion of Manuscript for Publication 112
References 147

List of Figures

4.1 let-23 – sevenless traversal 43

5.1 "Search all" using sis-b 54
5.2 Results of search for sis-b 55
5.3 Invoked thesaurus 55
5.4 Terms related to sis-b 56
5.5 Terms related to sex determination and signal 57
5.6 "Search all" using sdc-1 58
5.7 Results of search for sdc-1 58
5.8 Results of search for sdc-1 AND signal 59
5.9 ANOVA analysis for relevant documents 61
5.10 ANOVA analysis for recall 62
5.11 ANOVA analysis for precision 62

List of Tables

3.1 Number of overlapping terms between fly and worm thesauri 40

4.1 Number of nodes (whole phrases) found in conjoined fly-worm thesaurus 46

4.2 Number of nodes (partial phrases) found in conjoined fly-worm thesaurus 47

4.3 Number of suggested links found in conjoint thesaurus 49

C.1 Number of Iterations per New Search Used by Subjects While Browsing the Fly Thesaurus 101

C.2 Object Types for Terms at Various Traversal Positions 101

C.3 Number of intermediate nodes in traversal – entire phrase 102

C.4 Query terms found in concept space, by object type 102

C.5 Number of instances of various search heuristics using concept space . . 103

Chapter 1

INTRODUCTION

The Human Genome Initiative (HGI) offers tremendous challenges not only to the biology, biomedicine, and genetics research communities, but also to the information science and computer science communities. According to Courteau (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. The vast amount of knowledge accumulated during the project's scientific discovery process can only be managed with the use of computing technologies that support efficient and effective storage, retrieval, and analysis of information, that foster seamless distributed scientific collaboration, and that facilitate timely information sharing and effective information retrieval.

Biological research is highly data-intensive. 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. While the potential gains are undeniable, certain inherent problems run directly

counter to the highly collaborative nature of scientific investigation. Within the context of information sharing, scientific advancement is negatively affected by such problems as information overload, scattering of information, incompatibility of data representation between different databases, and vocabulary differences between subdisciplines.

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. The rate of growth for this already vast body of knowledge is estimated to be exponential (Frenkel, 1991). This *information overload problem* is further compounded by the parallel, distributed nature of biological research. 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 further complicates the problem of information access 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 that 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 under-

standing 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.

Information overload and the *vocabulary problem* in scientific research demand the development of advanced computing techniques. One recent attempt to address the problems of information overload and vocabulary differences in molecular biology research is the development of the Worm Community System (WCS) as part of the NSF *Collaboratory* effort (Rosenberg, 1992). 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 (Schatz, 1992) (Courteau, 1991). 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.

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 et al., ress). 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.

Chapter 1 provides an overview of the problem and the project. Chapter 2 present a literature review of cognitive studies and search aids that address the vocabulary problem. Chapter 3 describes a *concept space* approach, which is grounded on cluster analysis and general AI search algorithms, and provides a summary our previous findings. Chapter 4 reports the results of a cognitive study that investigated the concept (term) association behaviors of four biologists who are knowledgeable about both fly and worm genetics. Chapter 5 describes 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. Chapter 6 presents conclusions, a discussion of the contributions of this research, and anticipated future directions for the research. This thesis includes 5 appendices. Appendix A contains the instruments used in our system evaluation studies.. Appendix B contains a sample verbal protocol from each experiment conducted. Appendix C presents data resulting from analysis of verbal protocols and session logs. Appendix D includes a manuscript prepared for publication describing the generation and evaluation of the fly thesaurus.

Chapter 2

VOCABULARY ASSOCIATION AND VOCABULARY-BASED SEARCH AIDS

The problems of information overload and vocabulary difference affect every domain of human knowledge. Based on research over the past few decades, it has become clear to information scientists that development of online information retrieval systems must consider the preferences and cognitive processes of the users. In this section, we first look at some cognitive aspects of vocabulary association. We will then examine several automated approaches to easing the vocabulary problem in information retrieval.

2.1 Cognitive Aspects of Vocabulary Association

Cognitive studies have contributed greatly to the understanding of human intelligence and findings related to cognitive processes have significantly influenced the design of "intelligent" or knowledge-based systems (Newell and Simon, 1972) (Feigenbaum, 1977).

For the development of information retrieval systems, cognitive studies can be useful in two ways. First, by understanding how people perceive problems and their various approaches to problem solving, systems can be created that mimic those "intelligent" human processes. This is the essence of the General Problem Solver (GPS) in artificial intelligence (Newell and Simon, 1972) and is the foundation of expert or knowledge-based systems design (Hayes-Roth et al., 1983). Second, by studying how people use systems we can identify errors, misconceptions, and problems in those systems, and remedy them so as to improve system performance. This the realm of evaluation studies (Chen and Dhar, 1991) (Ramaprasad, 1987).

Both approaches to using cognitive studies have been adopted frequently in information retrieval research. While many studies have stressed search strategies (Ide and Salton, 1971) (McCall and Willett, 1986) (Chen and Dhar, 1991) and user modeling (Rich, 1983) (Daniels, 1986), this research focuses on the vocabulary association aspects of information retrieval.

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 et al., 1982a). Through the process of performing an information retrieval task, a searcher attempts to resolve those anomalies through a variety of retrieval strategies. Users' information needs are not, according to Belkin, "precisely specifiable." A user's vocabulary is often richer than is initially expressed in the problem statement. Palmquist and Balkrishnan (Palmquist and Balakrishnan, 1988) also found that a fuller vocabulary than that expressed in an initial problem statement can be elicited from the searcher during "continuous word association tests."

In Belkin's document retrieval system based on ASKs (Belkin et al., 1982a) (Belkin

et al., 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. As the searcher works through the task, "the anomaly and the user's perception of the problem will probably change with each instance of communication between the user and the system." Belkin concluded that information systems must be highly iterative and interactive (Belkin et al., 1982b). 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."

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) (Anderson, 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. Overlapping networks, those containing the same nodes, are interconnected parts of a larger network. Human long-term memory can be represented by such an interconnected network. Retrieval from long-term memory involves activation of a memory node. The act of remembering involves the spreading of activation along paths of associations from the activated node to the piece of information being sought. The strength of the association paths leading to that piece of information contributes to the level of activation being spread. Competing paths reduce the level of activation. This theory of *spreading activation* has influenced the design of many knowledge-based information retrieval systems (Shoval, 1981) (Cohen and Kjeldsen, 1987) (Chen and Ng, ress).

Anderson proposes that "indexing systems represent attempts to extend the organizing capabilities of the human mind to these artificial (humanly devised) information storage and communication systems" (Anderson, 1985b). Human memory and indexing systems perform three similar functions: they control the number and specificity of concepts; they relate terms, including synonyms, homonyms, and homographic equivalents, to concepts; and they connect associated concepts.

2.2 Vocabulary-based Search Aids

Based on mostly cognitive and user studies, query expansion is a well-known search strategy that searchers often use to improve on the effectiveness of the initial search. Hancock-Beaulieu (Hancock-Beaulieu, 1992) defines query expansion as "the process of adding new terms to a given query...for query formulation or reformulation." She observes that users "often change or add terms to an original query in the course of a search session."

Information retrieval in large document collections often requires query expansion aids because, as Blair and Maron (Blair and Maron, 1985) contend, "vocabulary problems make high recall impossible in full-text databases." Gomez et. al (Gomez et al., 1990) (Furnas et al., 1987) found in their studies that "searcher success is markedly improved by greatly increasing the number of names per object." Query expansion tools have been shown to be capable of improving retrieval performance significantly.

Many knowledge-based retrieval systems that perform query expansion have been created. For example, Hancock-Beaulieu (Hancock-Beaulieu, 1992) evaluated the effectiveness of an automatic query expansion facility included in the user interface of OKAPI, the online catalog at London's City University, over a six-month period. The tool adds to

the original query title terms, subject headings, and class numbers of records identified by the user as being relevant. She found that in searches where it was used, automatic query expansion accounted for 37% of all records chosen as relevant, and overall for 17% of all relevant records, including searches that were not expanded. Hancock-Beaulieu concluded that expanded searches were useful "to a substantial proportion of users."

Despite many positive results, numerous groups have reported poor results and even degraded performance with systems offering automatic query expansion, i.e., systems that automatically add terms to queries without the involvement of the user. Harman (Harman, 1988) showed performance degradation when terms from a statistically constructed thesaurus were added. However, when only those thesaurus terms which occurred in relevant documents were added, retrieval performance improved over that achieved by the original query. He suggested that the best retrieval performance may have been achieved when users filtered and selected candidate terms.

Several groups have created vocabulary-based search aids by making use of existing thesauri or dictionaries. While these tools are able to automatically provide the searcher with alternate terms to use in searching, they do not overcome the *knowledge acquisition bottleneck* (Hayes-Roth et al., 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*, which will be discussed in detail in the next section).

Fox et. al focused on creation of so-called "relational thesauri." For example, CODER adopted the *Handbook of Artificial Intelligence* and *Collin's Dictionary* (Fox, 1987) (Fox et al., 1988). Ahlswede and Evens parsed (Ahlswede and Evens, 1988) *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. In (Chen and Dhar, 1991), Chen and Dhar incorporated a portion of the *Library of Congress Subject Headings* (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 et al., 1993) (Chen and Ng, ress).

The National Library of Medicine's *Unified Medical Language System (UMLS)* project is probably the largest-scale effort 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 (McCray and Hole, 1990) (Lindberg and Humphreys, 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.

Chapter 3

THE CONCEPT SPACE APPROACH TO 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 functions) (Chen and Lynch, 1992) (Crouch, 1990) (Salton, 1989) (Rasmussen, 1992). Some algorithms, however, perform cluster analysis to further group terms of similar meanings (Everitt, 1980) (Rasmussen, 1992). Two assumptions underlie these techniques. First, terms that frequently occur together in documents are often about the same subject. This assumption appears to be consistent with how human beings articulate ideas and express knowledge (as chunks of similar concepts). Second, according to van Rijsbergen's *association hypothesis,* "if an index term is good at discriminating relevant from non-relevant documents, then any closely associated index term is also likely to be good at this" (Rijsbergen et al., 1981).

Stiles (Stiles, 1961) was one of the early researchers who reported improved retrieval

performance using a method based on term association (with collections of librarian-applied subject tags). Doyle (Doyle, 1962) further argued that the principles underlying association-based retrieval should apply whether the associations are determined by humans or by machines (programs). He suggested that, "if human associative processes are responsible for the associations found in text, then the reverse relationship should hold: that a human searcher, presented with a representation of text-derived associations, will be able to recognize as cognitive units many of the associated word pairs." Courtial and Pomian (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. "From the moment when the keywords indexing the documents of a database can be organized into associative networks, it is possible to breach the frontiers of their immediate domains of activity." 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), Lesk found little overlap existing between term relationships generated through term associations and those presented in existing thesauri. He purposely distinguished between "locally significant pairs" and those that are "significant in the absolute sense," the former making up the majority of term associations. He pointed out that these local pairs would not be useful for thesaurus construction, but "can point out word relations not normally apparent." He conceded, however, that "in larger collections, the apparent meanings of words may approximate their common meanings more closely." With regard to term relationships, the properties of "second-order associations" were also investigated. These are word pairs which need not co-occur in any documents, but rather have common first-order associations with a given term. We believe Lesk's concepts of collection completeness for capturing meanings of words and the second-order association properties are important but have often been overlooked in

prior research. (We will discuss this in more detail below.)

More recently, Crouch and Yang (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). Thesaurus "classes" are then formed from the low frequency terms of document clusters that pass a threshold value.

Ekmekcioglu et al. (Ekmekcioglu et al., 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 document 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, i.e., 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/she has generated.

3.1 Principles: The Concept Space Approach

Despite research over the past three 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 et al., 1972) (Peat and Willett, 1991) (Smeaton and van Rijsbergen, 1983). However, recall improvements of the order of 10 to 20 percent have been demonstrated when the thesaurus is used in an environment similar to that in which the original thesaurus was constructed (Salton, 1972) (Crouch, 1990) (Salton and Lesk, 1971). 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 requires a clear understanding of the following system development principles: *logarithmic vocabulary growth, completeness and recency, term specificity, asymmetric association, relevance feedback, vocabulary overlapping, and spreading activation.* Many of these principles are based on the human information processing theory (Newell and Simon, 1972) (Card et al., 1983) (Anderson, 1985a).

Based on these principles, 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) . We review these principles below in the context of our research:

- **Logarithmic vocabulary growth principle:**

  The most important rationale behind automatic thesaurus generation is related to the *information overload* problem. Lancaster has shown that the rate of growth for information (i.e. documents) continues at an exponential pace, while the corresponding rate of growth over the same period of time for number of concepts (keywords and terms) converges logarithmically (Lancaster, 1986). This principle appears to be applicable to different scientific domains and is particularly relevant in light of the rapid proliferation of Internet servers and distributed databases. We believe that, as *information space* continues to grow at such an alarming pace, the logarithmically converged *concept space* (network of domain-specific concepts and their associations) provides a manageable and understandable way to find (conceptually) relevant documents from a vast amount of online information.

- **Completeness and recency principle:**

  As we have discussed earlier, early information science researchers such as Lesk have suggested the importance of large document collections for generating automatic thesaurus (Lesk, 1969). This is especially true when considering the logarithmic vocabulary growth principle described above. If a collection which is used to generate an automatic thesaurus is limited, it is impossible to reach the plateau (or convergence) on the logarithmic curve. Many previous automatic thesaurus generation studies which used only selected collections and/or samples of documents in a subject domain suffered from a lack of completeness in document collections.

  In addition to the size consideration, researchers need to be aware that completeness is subject dependent. For a small subject domain (such as the worm biology, which has a short history and limited number of researchers (Chen et al., ress)), several thousand documents could be considered "complete" in light of all the publications

generated in so limited a domain. However, for a more general and large subject area (such as computer engineering (Chen and Schatz, 1994)), creating a complete domain-specific thesaurus may require using millions of documents.

For many scientific domains, the *fluidity* of concepts and vocabularies places a special burden on thesaurus builders (Frenkel, 1991). A manual approach to scientific thesaurus building often fails due to the lack of specificity and timeliness in representing new and emerging scientific concepts and knowledge. For automatic thesaurus generation, special emphasis should be placed on identifying complete recent collections. Although some scientists may use a document collection from a historical perspective (and thus are interested in old documents and using old vocabularies), the majority of the scientists (such as the molecular biologists) are likely to be more interested in finding new and ongoing research in their subject areas. When a scientific subject domain is too vast to identify a complete collection for automatic thesaurus generation, recent documents in the subject area may become more important for purposes of scientific collaboration and information sharing.

As computing resources become more abundant, we believe sampling of documents for automatic thesaurus generation is unnecessary and that collecting a complete and recent document set from existing online sources should be the first step toward creating a robust and useful domain-specific thesaurus.

- **Term specificity principle:**

  Most prior automatic thesaurus generation studies relied on automatic indexing techniques (Salton, 1989) (Ekmekcioglu et al., 1992). Words were identified, stemmed, and combined to produce descriptors (automatic indexes of single or multiple keywords) to represent the content of a document. The process is simple

and domain-independent, and does not require the extensive linguistic or domain-specific knowledge that is often essential for artificial intelligence based natural language processing (NLP) (Woods, 1972) (Fillmore, 1968) (Burton, 1976).

Despite its simplicity, automatic indexing may produce significant amounts of "noise," e.g., typos, meaningless acronyms, general terms, and random permutation of adjacent terms. Special attention needs to be paid to generate specific and meaningful terms and a combination of techniques needs to be used. *Term frequency* and *inverse document frequency* are required to reward terms that are specific. Empirical thresholds (of term occurrences) should be adopted to remove terms that are incidental typos or random term permutations. Modification of stemming algorithms need to be performed to accommodate the special characteristics of applications (for example, cloning and clones are two related but different biological concepts, thus should not be stemmed to the same root form, clone).

One important observation from our past research has been appreciation of the abundance and availability of existing vocabulary lists and their importance for identifying specific and useful domain-specific concepts. For most subject domains, subject indexes and researcher names often can be obtained easily by scanning the entries at the back of textbooks or technical manuscripts. The white page service and various distributed sources on Internet also make available lists of domain-specific vocabularies such as subject terms, researcher names, company names, gene names, experimental methods, and so on. In almost every application domain that we have studied, including Russian computing (Chen and Lynch, 1992), business (Chen et al., 1994a), worm biology (Chen et al., ress), fly biology (Chen et al., 1994b), engineering (Chen and Schatz, 1994), we have found extensive vocabulary lists that can help identify specific content descriptors in a document. (We refer to this process as *object filtering* in our research.) By applying the

combination of object filtering and careful automatic indexing, we found that the resulting automatic thesauri were often less noisy and more useful for users of the specific domains.

- **Asymmetric association principle:**

  Human memory association is asymmetric by nature (Anderson, 1985a). That is, the strength of the association from term A to term B is often different from the strength of association from term B to term A (for example, it is much easier to associate "net" with "volleyball" than "volleyball" with "net"). This characteristic is also evident in scientific information retrieval. For example, when looking for work related to a researcher named "Mary Smith," a searcher is likely to find articles specifically related to her investigations in the area of "embryonic development;" but when performing a search using "embryonic development," it is unlikely that documents related to "Mary Smith" will be retrieved because a multitude of biologists study embryonic development (an example from our recent experiment in molecular biology).

  However, this asymmetric association property of human memory and information retrieval had not been considered in most prevailing similarity functions. The limitation of the popular symmetric similarity functions, e.g., cosine, Dice, and Jaccard's, have been reported recently by Peat and Willett (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 (related

to the *specificity* principle discussed above).

We echo their observations and, in fact, we have independently reached the same conclusion through our experience in developing automatic thesauri and working with domain experts in several applications (the thesauri generated in our research were often the results of iterative prototyping and refinement based on cognitive studies performed with the domain experts and their suggestions). An asymmetric similarity function was developed and shown to be more accurate in representing human memory association than the popular cosine function (Chen and Lynch, 1992). More recently, further modification was made to reward related terms that are specific (a revised *inverse document frequency* function was used to reward specific terms and penalize general terms during co-occurrence analysis) (Chen et al., ress). This asymmetric similarity function has since become an integral part of our proposed *concept space* approach to automatic thesaurus generation.

In some system development efforts, more elaborate hierarchical clustering methods (Rasmussen, 1992) (Crouch and Yang, 1992) and/or neural network grouping techniques (Chen et al., 1994a) may be performed following co-occurrence analysis to produce clusters of similar terms. While this type of analysis is appropriate for abstraction and categorization purposes, the outputs often lack the intuitive association between concepts and are incomprehensible to users seeking to expand a query. In our experience, weighted, asymmetric networks of concept pairs, akin to human-made thesauri, appear to be more intuitive and easier for searchers of varying backgrounds to use.

- **Relevance feedback principle:**

  Harman (Harman, 1988) suggested that the best retrieval performance was achieved when users filtered and selected candidate terms. Croft and Das (Croft and Das, 1990) also reported significant improvements in effectiveness of expanded queries when users are prompted for additional terms that can be used in the search. Automatic term replacement or switching is often misleading and impractical, considering the unique context and backgrounds that different searchers might have. We believe an interactive relevance-feedback process of term selection is essential to the effective usage of automatic thesaurus. (Human information specialists have been observed to perform extensive user modeling and query articulation when assisting patrons in using an existing thesaurus (Chen and Dhar, 1987).)

- **Vocabulary overlapping principle:**

  Numerous investigators in information retrieval have suggested the idea of "switching" languages, which could be consulted either automatically or manually, to aid searchers in performing multiple-database searches. Lancaster, in discussing compatibility and convertibility of vocabularies between databases, contended that because controlled vocabularies tended to promote internal consistency within individual databases and information systems, they often reduced compatibility between systems (Lancaster, 1986). Lancaster suggested that a "neutral switching language" can be used to convert from any one vocabulary into another. While he was clearly referring to a manually developed switching language, his notion of an "intermediate lexicon" is the conceptual basis of our approach to bridging the vocabulary differences between different scientific domains.

  For scientific collaboration and information sharing across different domains, multiple domain-specific thesauri (existing or automatically generated) need to be created

and coupled in order to assist in cross-domain term switching. The overlapping vocabularies in different, but somewhat related domains (e.g., fly biology and worm biology, computer science and electrical engineering) create potential vocabulary paths from one domain to the other, which can bridge the vocabulary differences during information retrieval.

The vocabulary overlapping principle is also the rationale behind the National Library of Medicine's UMLS project for automatically suggesting biomedical terms for different databases (McCray and Hole, 1990) (Lindberg and Humphreys, 1990). 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 and Ng, ress). In order to assist in cross-domain scientific information sharing, we believe we need to create a domain-specific thesaurus for the underlying database (e.g., a worm thesaurus for the Worm Community System) and several related thesauri for other community outsiders (e.g., fly thesaurus, rat thesaurus, e. coli thesaurus, etc.). These *concept spaces* will overlap and provide vocabulary paths for supporting cross-domain information retrieval. This paper will report in detail an experiment which aimed to assist fly biologists in retrieving worm documents using a conjoined fly-worm thesaurus.

- **Spreading activation principle:**

  During the course of designing several large-scale, domain-specific thesauri, we found that the most frequent complaints from users who performed term switching manually (i.e., in a user-controlled browsing mode) was that the process was tedious and cognitively demanding and that users often got lost after exploring a number of concepts. These browsing problems are not unfamiliar to developers of large-scale

hypertext systems, e.g., the *embedded digression problem* (a system may confuse and disorient its user) and the *art museum phenomenon* (a system could cause the user to spend a great deal of time while learning nothing specific) (Foss, 1989) (Carmel et al., 1992).

Causes of such problems may be related to the second-order association described by Lesk (Lesk, 1969). Some terms may be related indirectly via their common associations with another term. Humans often perform such multiple-link associations (e.g., A is related to B, which in turn is related to C; C is related to both A and B) during problem solving or long-term memory recall, a process frequently referred to as *spreading activation* (Anderson, 1985a). Both Kim and Kim (Kim and Kim, 1990) and Chen et. al (Chen et al., 1993) proposed treating a thesaurus as a neural network or semantic network and applying spreading activation algorithms. Despite the lack of published research that supports the usefulness of spreading activation algorithms for term suggestion (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 and Ng, ress). Our proposed algorithmic approach to associative retrieval appeared to be a viable option for efficiently traversing large-scale, multiple thesauri across different domains.

3.2 Techniques: The Concept Space Approach

Based on the seven principles described above, we developed selected algorithms for automatic thesaurus generation. We believe our *concept space* approach, if applied properly, can be extremely powerful in capturing the domain knowledge in textual databases and creating an environment for concept-based information management and retrieval. The specific steps and algorithms adopted include: *document and object list collection*, *object filtering and automatic indexing*, *cluster 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 and Lynch, 1992) (Chen et al., 1993) (Chen et al., 1994a) (Chen et al., ress) (Chen et al., 1994b) (Chen and Ng, ress).

- **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.

  In (Bates, 1986), Bates 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 through the helps 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 followed. In (Salton, 1989), Salton 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., ress), we have made several changes to 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.

- **Cluster 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 fre-

quency, 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:

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

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.

We then performed term co-occurrence analysis based on the asymmetric "Cluster Function" developed by Chen and Lynch (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.

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

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

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:

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

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

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.

- **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 and Ng, ress). The Hopfield algorithm, in particular, has been shown to be ideal for concept-based information retrieval.

  The Hopfield net (Hopfield, 1982) was introduced as a neural network that can be used as a content-addressable memory. Knowledge and information can be stored in single-layered, interconnected 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 (Tank and Hopfield, 1987) (Knight, 1990) 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*, where 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

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

$\mu_j(t+1)$ 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 (Knight, 1990) (Dalton and Deshmane, 1991). This formula shows the *parallel relaxation* property of the Hopfield net. (Readers are referred to (Chen and Ng, ress) 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 and Schatz, 1994).

3.3 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 two years, we generated a worm thesaurus in Fall 1993 (Chen et al., ress) 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 at: http://bpaosf.bpa.arizona.edu:8000/cgi-bin/BioQuest.

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 3.3). 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 GUI interface) on the operational Worm Community

Objects	Fly		Overlap	Worm	
	Total fly terms	% Overlapping with worm	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

Table 3.1: Number of overlapping terms between fly and worm thesauri

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).

A manuscript describing the generation and evaluation of the fly thesaurus is appended to this thesis (Appendix E). In that study, the author was responsible for creating the document collections underlying thesaurus (FlyBase files, and Medline and Biosis records), and for conducting and analyzing the results of the thesaurus evaluation experiments, including development of the taxonomy of system problems.

Chapter 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 worm – 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 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 five 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 5 traversals, to 1 hour 30 minutes for one graduate student who completed 3 traversals.

4.2 A Sample Traversal and Analysis of Traversal Graphs

Figure 4.1 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

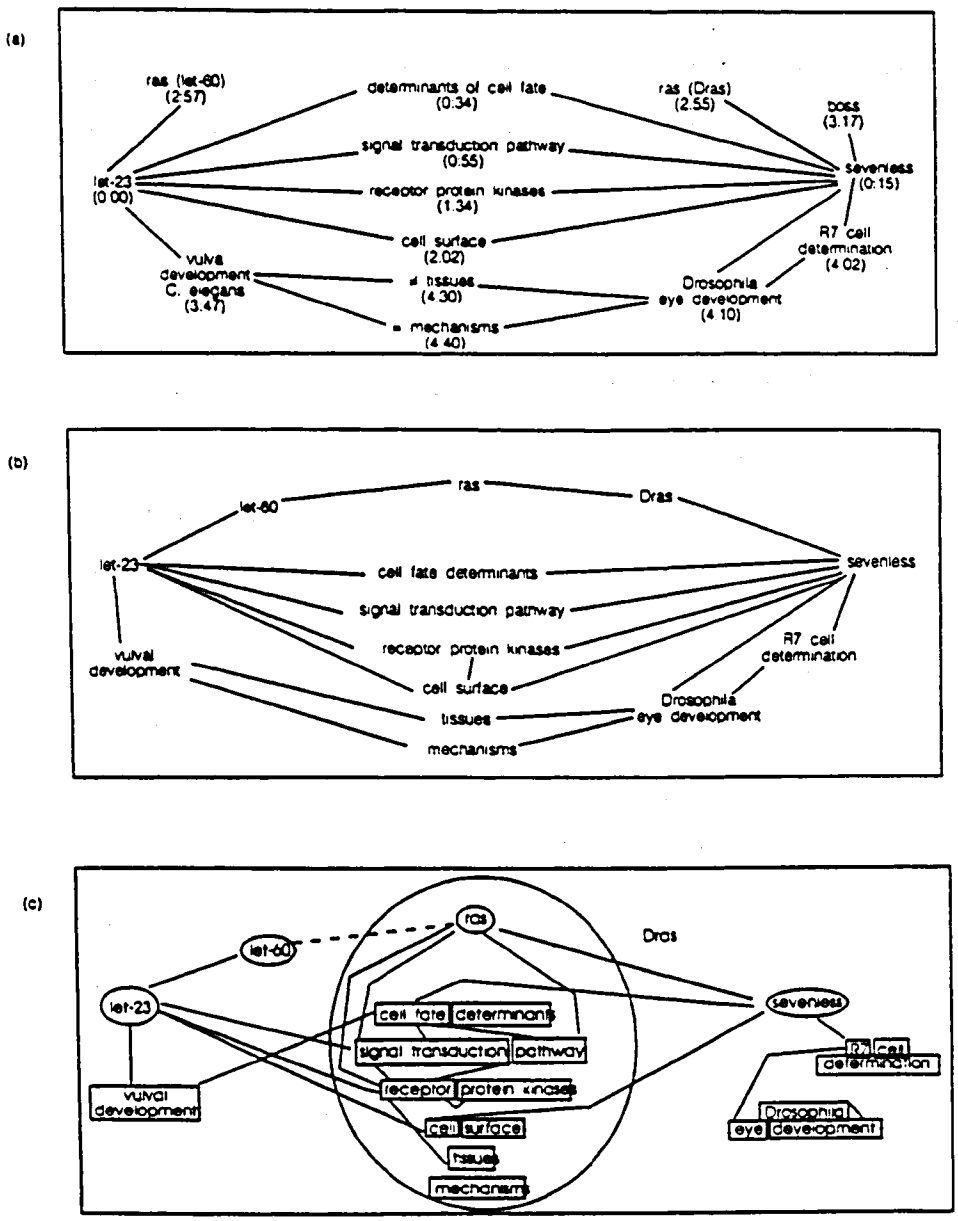


Figure 4.1: let-23 – sevenless traversal

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 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 4.1, components of all term phrases were found, however 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:

- **A large percentage of the worm and fly terms were found in the thesaurus:**

  Tables 4.1 and 4.2 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.

Subject	Object	Worm Specific Found/Suggested	Fly Specific Found/Suggested	Common Term Found/Suggested	Total Found/Suggested	Percent Found
	Author	2/2	1/1	0/0	3/3	100%
Worm	Function	0/0	0/0	0/0	0/0	
Experts	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%	
	Author	0/0	0/0	0/0	0/0	
Fly	Function	0/0	0/0	0/0	0/0	
Experts	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%	62.5%	45.0%	60.0%	
	Author	2/2	1/1	0/0	3/3	100%
	Function	0/0	0/0	0/0	0/0	
Overall	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%	

Table 4.1: Number of nodes (whole phrases) found in conjoined fly-worm thesaurus

Subject	Object	Worm Found/Suggested	Fly Found/Suggested	Common Found/Suggested	Total Found/Suggested	Percent Found
	Author	2/2	1/1	0/0	3/3	100%
Worm	Function	0/0	0/0	2/2	2/2	100%
Experts	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%	
	Author	0/0	0/0	0/0	0/0	
Fly	Function	0/0	0/0	0/0	0/0	
Experts	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%	90.0%	89.7%	85.6%	
	Author	2/2	1/1	0/0	3/3	100%
	Function	0/0	0/0	2/2	2/2	100%
Overall	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 4.2: Number of nodes (partial phrases) found in conjoined fly-worm thesaurus

- **A small percentage of the term associations were found in the thesaurus:**

  Figure 4.3 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 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 Behavior

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

Subject	Link	Partial Phrases		Whole Phrases	
		Subject Suggested	Additional Found	Subject Suggested	Additional Found
	fly - fly	2/7	0	0/7	3
Worm	fly - common	6/56	2	2/33	0
Experts	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 - fly	3/10	1	2/6	0
Fly	fly - common	15/67	4	3/38	0
Experts	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	15/182	21
	Percent	18.7%		17/199	
	fly - fly	5/17	1	2/13	3
	fly - common	21/123	6	5/71	0
Overall	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%	

Table 4.3: Number of suggested links found in conjoint thesaurus

- **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 associations:**

  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 terminology (Bates, 1986)). However, the association links suggested by the thesaurus were often different from those provided by the subjects. The usefulness of the thesaurus associations needed to be investigated, especially for cross-domain scientific information retrieval.

Chapter 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. A simple GUI interface that was incorporated allowed subjects to browse the thesaurus. 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 performed searches using the Worm Community System, with the aim of identifying

useful or relevant worm documents in response to a Drosophila-related query. 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 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 one and two hours for their queries. The super-expert, a Drosophila researcher (faculty) with over 10 years of experience in the field, spent almost ten hours in reviewing all the results.

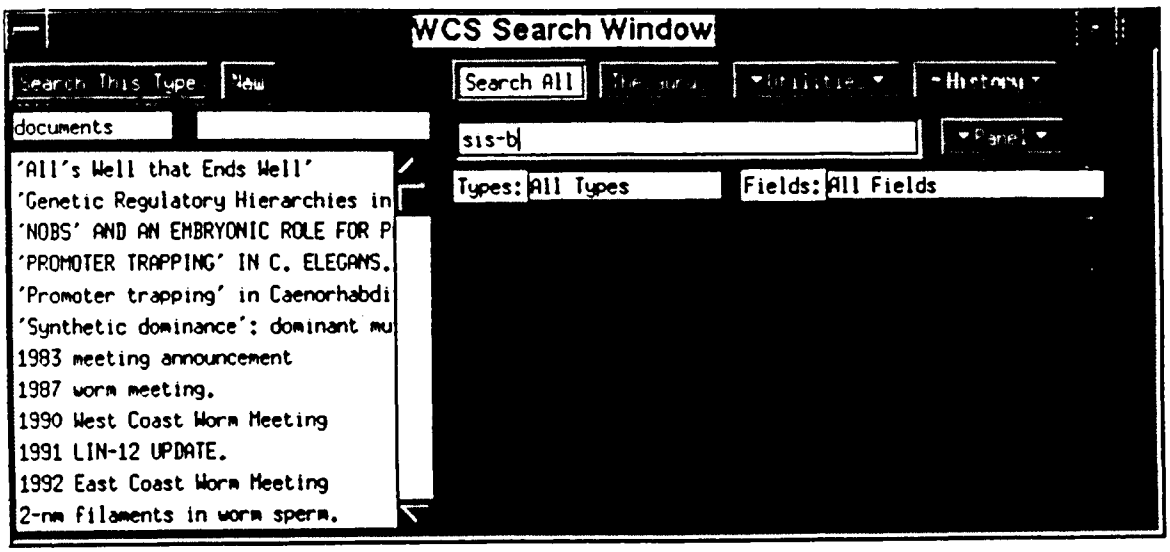


Figure 5.1: "Search all" using sis-b

5.2 A WCS Sample Search

A sequence of query operations for worm documents related to a gene known in the fly domain as sisterless-b (abbreviated as sis-b) is shown in Figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, and 5.8. A fly researcher approaching the Worm Community System entered the fly gene name "sis-b," and clicked on the "Search All" button (highlighted in white, Figure 5.1) to activate a search of the worm database. Moments later another window was displayed containing the search results (Figure 5.2). Four items, all documents (as indicated by the object type indicator DOC), were retrieved. The user looked for clues as to the relevance of the items by reading the titles and/or full text of the document (the items can be "opened" by double clicking on the title). Convinced that none satisfied the query, the user went back to the "WCS Search Window" and activated the thesaurus (highlighted in white, Figure 5.3). Soon, the Thesaurus Window appeared (Figure 5.4).

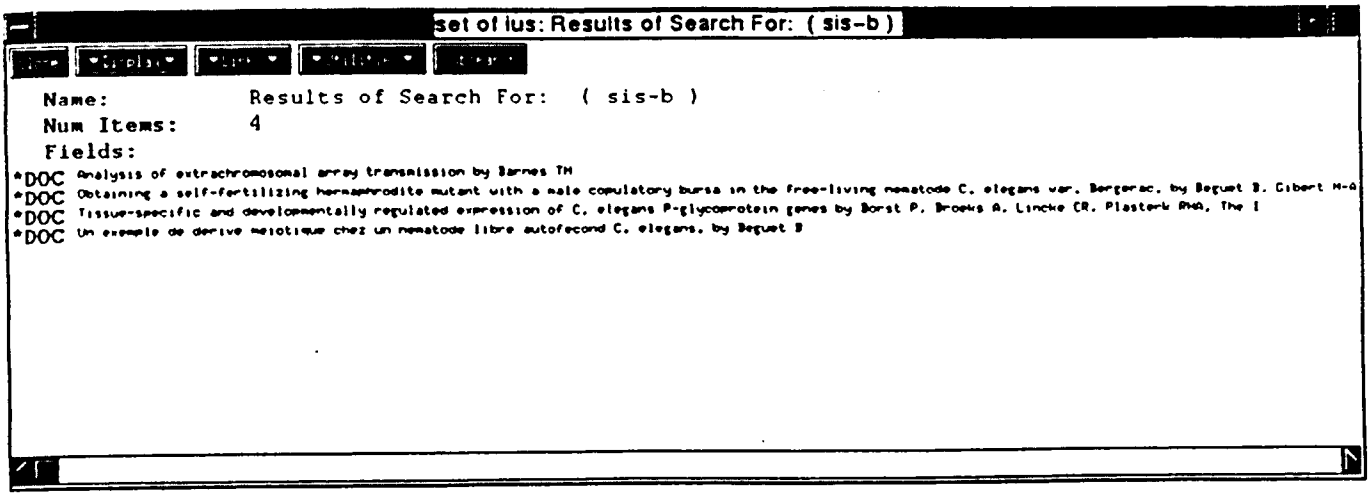


Figure 5.2: Results of search for sis-b

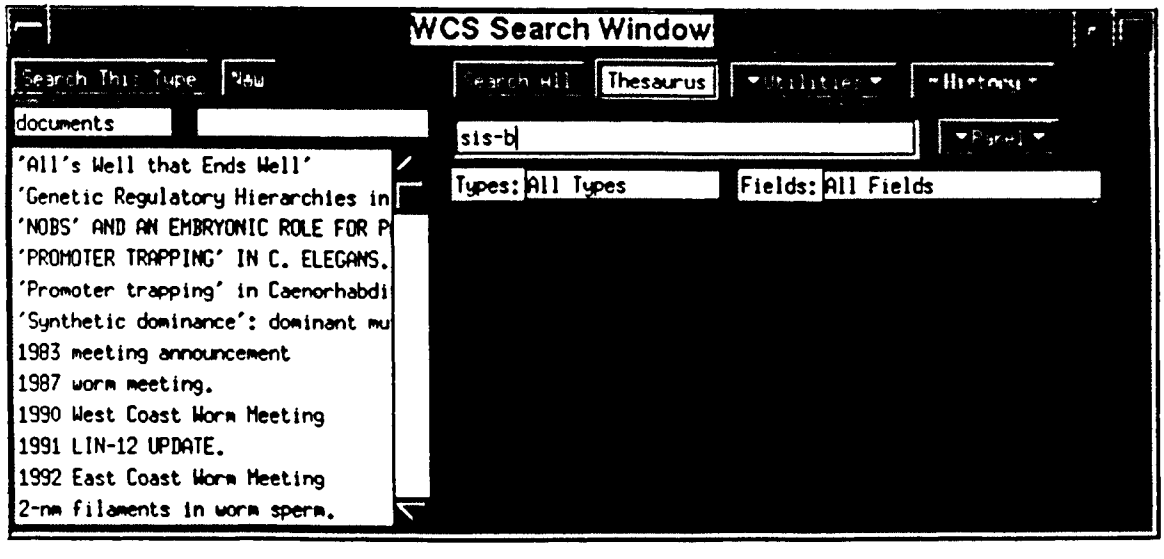


Figure 5.3: Invoked thesaurus

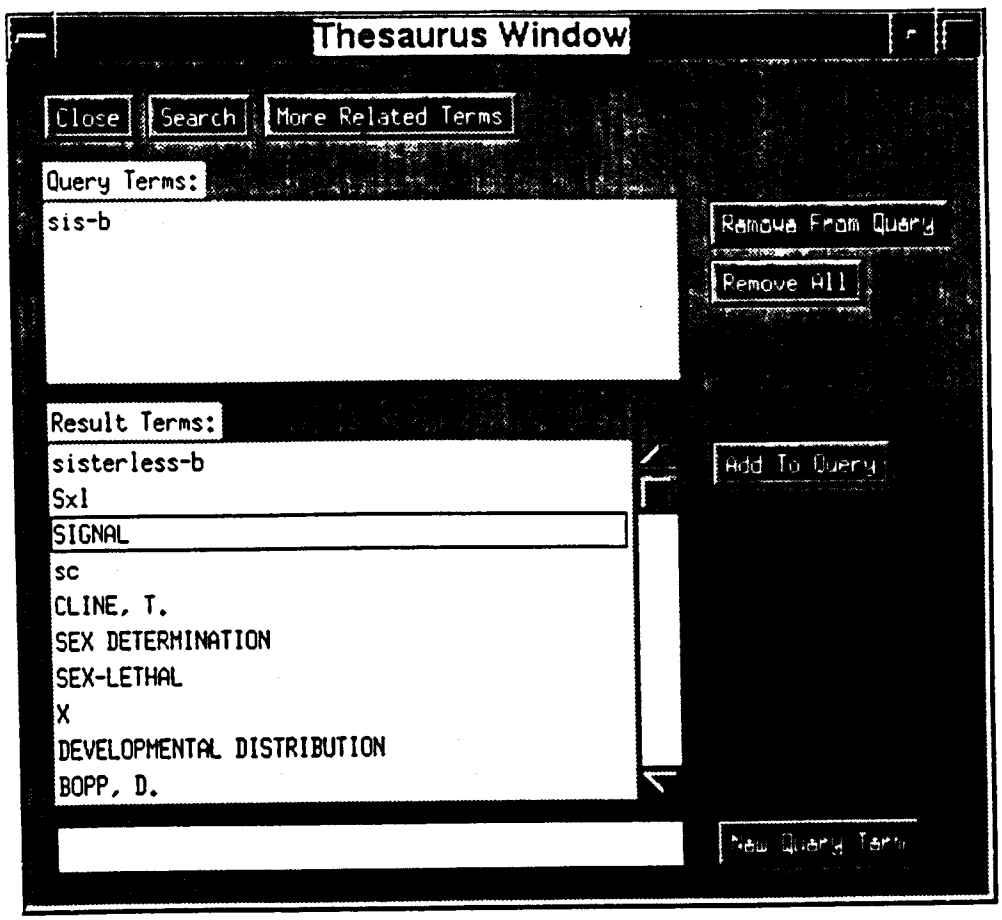


Figure 5.4: Terms related to sis-b

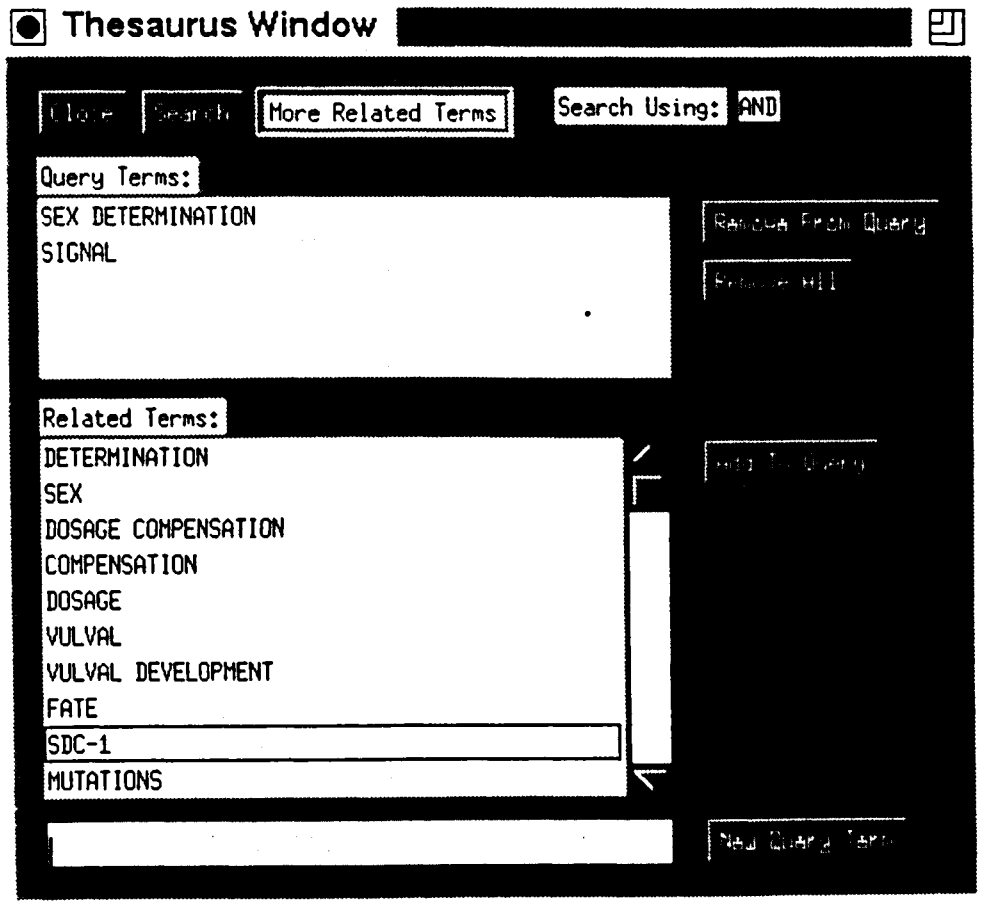


Figure 5.5: Terms related to sex determination and signal

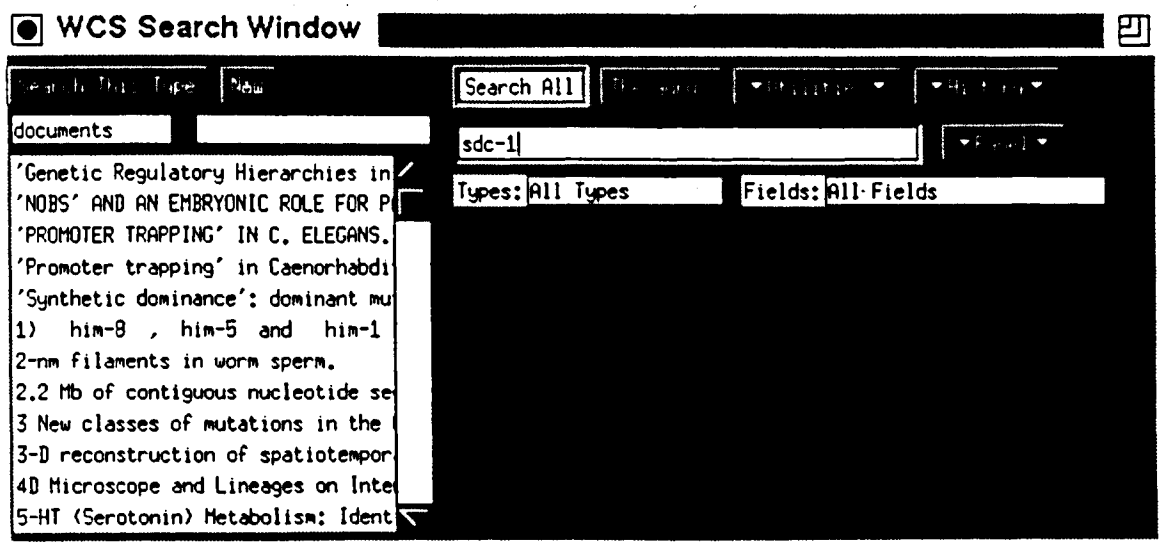


Figure 5.6: "Search all" using sdc-1

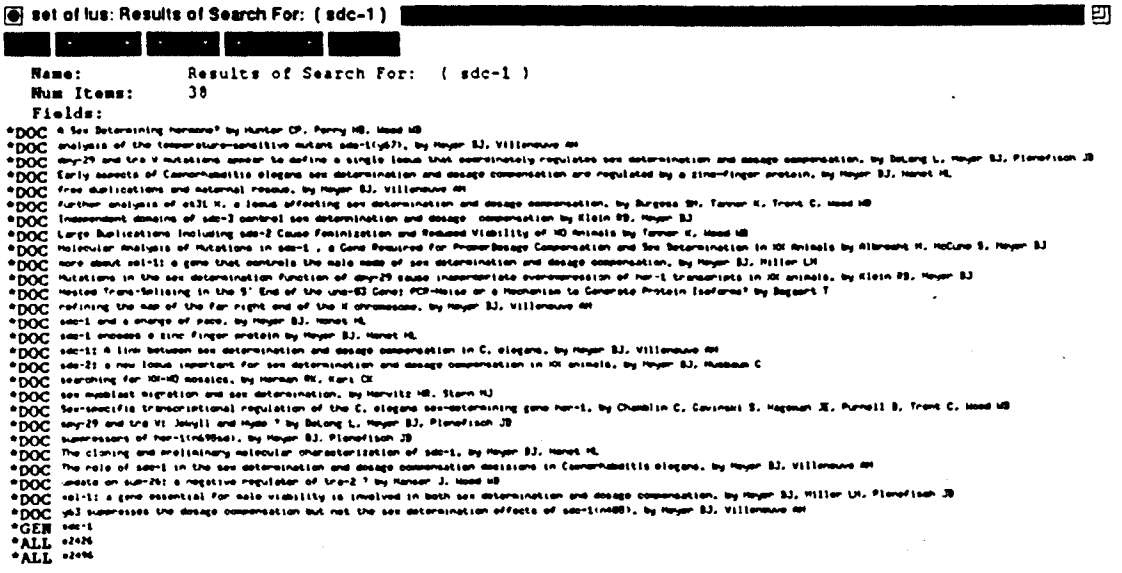


Figure 5.7: Results of search for sdc-1

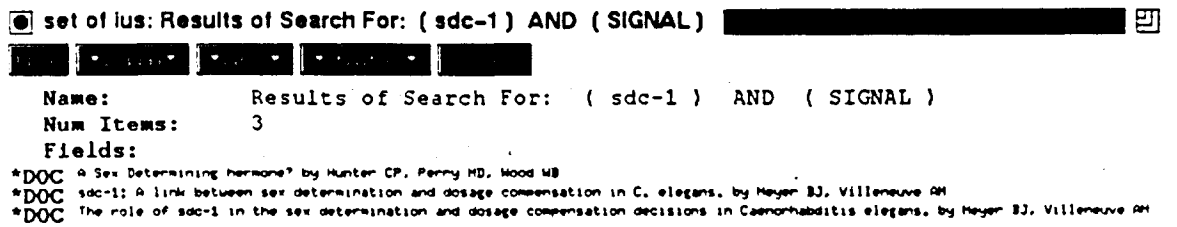


Figure 5.8: Results of search for sdc-1 AND signal

The query term "sis-b" appeared in the "Query Terms" box and related terms were displayed in the "Result Terms" box. High on the list were "sisterless-b", the full name of the gene, and two very closely related fly genes (sc and Sxl). The names of two researchers (Cline, T. and Bopp, D.), and several terms indicative of the gene's function: "SIGNAL," "SEX DETERMINATION," "SEX-LETHAL," "X" (the X chromosome), and "DEVELOPMENTAL DISTRIBUTION" also appeared. The user, stating an interest in finding documents about worm genes involved in sex determination that function in signaling pathways, highlighted "SEX DETERMINATION" and "SIGNAL" and clicked on "Add to Query" to add each to the "Query Terms" box. The user removed "sis-b" from the list using the "Remove from Query" button to the right of the "Query Terms" box, then clicked on the "More Related Terms" button at the top of the interface. On the new list of system-suggested terms (Figure 5.5), the user found the gene name "SDC-1." The user stated an interest in finding documents about the role of sdc-1 as a signaling

protein and entered "sdc-1" in the WCS Search window (Figure 5.6), which retrieved 38 documents (Figure 5.7). Considering this too large a set, the user added "SIGNAL" to the query, using the Boolean operator "and." This retrieved a set of three documents, all of which the user found relevant (Figure 5.8).

5.3 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. Results of this experiment are summarized below:

- **The conjoined thesaurus helped find more relevant documents:**

  Results from calculation of relevant documents retrieved, presented in Figure 5.9, 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 duplicate much with the set of documents retrieved without the use of the thesaurus. One-way analysis of variance (ANOVA) using the MINITAB statistical package (Ryan et al., 1985) showed that this improvement was

```
ANALYSIS OF VARIANCE
SOURCE      DF        SS        MS        F        p
FACTOR       1       869       869     3.89    0.059
ERROR       26      5803       223
TOTAL       27      6672

                                        INDIVIDUAL 95 PCT CI'S FOR MEAN
                                        BASED ON POOLED STDEV
LEVEL               N     MEAN     STDEV  ----------+---------+---------+------
Docs without       14     8.79     10.61   (---------*---------)
Docs with          14    19.93     18.27                (---------*---------)
                                          ----------+---------+---------+------
POOLED STDEV =    14.94                             8.0       16.0      24.0
```

Figure 5.9: ANOVA analysis for relevant documents

statistically significant (P=0.059). (In all our analysis, a 10% statistical significance level was adopted.)

- **The conjoined thesaurus helped improve 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. As shown in Figure 5.10, 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:**

  Figure 5.11 shows the results of precision computation for all subjects. The overall precision was 43.51%[1] without the thesaurus and 53.48% with the thesaurus. However, this improvement was not statistically significant (P=0.477).

[1]The sample size for precision computation was 13 instead of 14 because one subject did not identify any relevant documents during his initial search without the thesaurus.

```
ANALYSIS OF VARIANCE
SOURCE      DF        SS        MS        F        p
FACTOR       1     0.756     0.756     6.72    0.015
ERROR       26     2.925     0.112
TOTAL       27     3.681

                                          INDIVIDUAL 95 PCT CI'S FOR MEAN
                                          BASED ON POOLED STDEV
LEVEL              N      MEAN     STDEV  ----+---------+---------+---------+--
Recall without    14    0.3241    0.3391   (--------*--------)
Recall with       14    0.6528    0.3316                     (---------*--------)
                                          ----+---------+---------+---------+--
POOLED STDEV =  0.3354                      0.20      0.40      0.60      0.80
```

Figure 5.10: ANOVA analysis for recall

```
ANALYSIS OF VARIANCE
SOURCE      DF        SS        MS        F        p
FACTOR       1     0.067     0.067     0.52    0.477
ERROR       25     3.221     0.129
TOTAL       26     3.288

                                          INDIVIDUAL 95 PCT CI'S FOR MEAN
                                          BASED ON POOLED STDEV
LEVEL              N      MEAN     STDEV  -----+---------+---------+---------+-
Precision w/o     13    0.4351    0.3845  (-------------*-------------)
Precision with    14    0.5348    0.3336        (-------------*-------------)
                                          -----+---------+---------+---------+-
POOLED STDEV =  0.3590                      0.30      0.45      0.60      0.75
```

Figure 5.11: ANOVA analysis for precision

5.4 Experimental Results: Search Behavior

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 holds different meaning 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 documents 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 toward learning what is known about a particular biological system in the worm or toward 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 don't exist in worms, e.g., genes or proteins related to wing function (the fly has wings, but not the worm).

- **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." This sort of trial-and-error approach, resulting from serendipitous discovery, while educational, was not very efficient.

- **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.

Chapter 6

CONCLUSIONS

Information overload and the *vocabulary problem* in scientific research demand the development of advanced computing techniques. This paper has presented a *concept space* approach to addressing the vocabulary problem in scientific collaboration and information sharing, using molecular biology domain as an example. We first provide a literature review of cognitive studies related to the vocabulary problem and vocabulary-based search aids first. 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* was then described to provide a theoretical foundation for query expansion in information retrieval.

Despite many positive results, numerous groups have also 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 seven important principles for automatic thesaurus generation: *logarithmic vocabulary growth, completeness and recency, term specificity, asymmetric association, relevance feedback, vocabulary overlapping, and spreading acti-*

vation. The specific steps and algorithms adopted in our *concept space* approach include: *document and object list collection, object filtering and automatic indexing, cluster analysis,* and *associative retrieval.*

In an attempt to understand the usefulness and performance 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 we only found a small percentage (8.4%-18.1%) of the associations suggested by the subjects 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* on the Internet WWW for remote access (http://bpaosf.bpa.arizona.edu:8000/cgi-bin/BioQuest). *BioQuest* contains several thousand documents in worm biology and allows WAIS-like key-

word 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.

Appendix A

Experimental Instruments

Fly Thesaurus Evaluation Experiment: Subject Briefing Statement

Purpose

The purpose of this experiment is to determine whether certain fly biology terms are related. You will be asked to generate associa- tions to ten selected terms, then to review the terms the system suggests. Lastly, you will be asked to evaluate the system and give feedback about its accuracy. The complete experiment will last for about an hour.

Experiment

Part I: Term Association

You are asked to associate as many terms as possible with the ten terms.

Part II: Evaluation of Thesaurus Associations

Please evaluate the term associations suggested by the Fly thesaurus.

Part III: Thesaurus Browsing

Please browse the generated Fly Thesaurus freely, think aloud during this process, and give any suggestions/observations that you have about the relevance of the concept associations.

This part of the experiment utilizes a procedure called Protocol Analysis. This analysis requires subject articulation. The subject is asked to speak aloud during the experiment in order to understand how the subject associates concepts. There are three aspects to this procedure:

1) Your are asked to think aloud during all aspects of the experiment.

2) The entire experiment will be recorded on tape to capture the verbal expressions and further understand your term associations.

3) Experimenters will be present at the experiment taking notes of actions performed.

Part IV: Questions

1) Do you have any comments, observations, or suggestions regarding the quality of the Fly Thesaurus?

2) Do you have any comments, observations, or suggestions regarding the user interface?

Thank you for your participation.

A.1 Fly-Worm Traversal Experiment: Subject Briefing Statement

Purpose

The goal of this experiment is to understand your associations between concepts that form the basis for the decisions and the inferences that you make when searching information.

We would like to you to identify some paths of associated terms that might be taken to traverse from terms in one domain to terms in the other domain that are known by different names in fly and worm.

Experiment

First, you will be asked to identify pairs of terms – one in fly and one in worm – that you know to be related. For each pair, you will develop a path from one term to the other, using other related terms. You are asked to articulate clearly any thoughts that occur to you as you identify these associations, particularly those that are descriptive of how you arrived at an associated term.

We will go through one traversal task together so that you get a feel for the kind of knowledge we would like to capture. The complete experiment will last about an hour.

There are three aspects to this procedure:

1) Subjects are asked to think aloud during all aspects of the experiment.

2) The entire experiment will be recorded on tape to capture the verbal expressions and further understand the subject's linkages.

3) I will be taking some notes during the experiment to help in analysis of your verbal protocols.

Thank you for your participation.

A.2 Fly-Worm-WCS Document Retrieval Experiment: Subject Briefing Statement

Purpose

The Worm Community System is a digital library that contains knowledge about C. elegans. The system permits searching and browsing of the existing knowledge in the research community, as well as addition of data and literature from users in remote sites.

The Fly-Worm Thesaurus Traversal project is based on the premise that knowledge gained from progress in one domain of biology may be useful to researchers in other biological domains, but that the differences in vocabulary between those two domains will preclude access to and retrieval from the literature of the other domain. The goal of this experiment is to determine the effectiveness of the system we have created for aiding retrieval relevant worm documents in response to a fly-specific query.

Experiment

Part I: WCS Document Retrieval

You are asked to perform 3-4 searches using fly terms. There will be two steps to each query conducted. First, the fly term will be searched in the WCS, and you will be asked to determine the relevance of items in each set of retrieved objects within the context of the query statement.

Second, you will use the conjoint Fly-Worm Thesaurus to identify terms to add to your initial query and improve the search results. Again, you will be asked to determine the relevance of the retrieved items.

In both steps, the full capabilities of the WCS will be at your disposal. Boolean

searching is available, as is access to such other objects, such as genes, alleles, contigs, deficiencies, sequences, rearrangements, chromosomes, strains, persons, genetic crosses, cell lineages, alleles, etc.

You are asked to articulate clearly as you evaluate the outcome of each search. Your verbal protocol will be tape recorded for later analysis. The entire experiment will last about an hour and a half.

Part II: Questions

1) Do you have any comments, observations, or suggestions regarding the quality of the Fly Thesaurus?

2) Do you have any comments, observations, or suggestions regarding the user interface?

Thank you for your participation.

A.3 Fly-Worm-WCS Document Retrieval Experiment:Subject Queries Presented to Super-expert

Query 1 – I am interesting in knowing if worm biologists have found a homolog of wingless in worms, and what they know about it.

Query 2 – We know that the posterior group genes in flies are cytoplasmic determinants that play a role in gametogenesis. I want to know what mutants have been found in worms that participate in making P granules.

Query 3 – I want to know what people have found in worms about the autonomy of

the EGF receptor.

Query 4 – I know that cytogenetics is very big in flies, but I don't know if they study it in worms. In particular, I am interested in finding out what they know about position effect variegation.

Query 5 – I have been thinking a lot about neomorphic and antimorphic alleles. I'm curious about what they think about the concept.

Query 6 – Mike Levine does work on spatial development in the embryo. I would like to know what similar work is being done by worm people, especially with regard to transcription of homeobox genes.

Query 7 – Broad Complex is a complex genetic locus that codes for transcription factors that are activated during metamorphosis by steroid hormones. I would to know if there are any genetic or molecular equivalents or homologs in the worm field.

Query 8 – Troponin is a muscle protein in Drosophila that regulates contraction. There are several forms of it in Drosophila. In our lab, we have found Troponin I and Troponin T. I'm interested in seeing what they know in worms.

Query 9 – We're looking at chromatin structure, and the relationship between gene expression and structure. I'd like to know what kinds of studies they are doing in worms with regard to maintaining chromatin structure.

Query 10 – I'd like to know what worm biologists know about neurogenic genes that are expressed during embryogenesis.

Query 11 – I use a technique called in situ hybridization to study transcription of embryonic genes. I want to know what kinds of approaches they take in using this technique. One example might be using biotin in situ hybridization to study chromosomes.

Query 12 – I'd like to know how worm biologists perform fixation of microtubules.

Query 13 – I'd like to what they know about the structure of meiotic chromosomes.

Query 14 – I'd like to know what is the worm homolog for sevenless.

Appendix B

Sample Verbal Protocols

Included in this appendix are sample verbal protocols, one from each of the three experiments conducted, to give a flavor of each experiment.

Fly Thesaurus Evaluation Experiment

Subject 4 – Outsider

Exp: If you see one here that looks interesting to you, you can mark it. And just use it as if you were searching.

Subj: I hate computers. Sorry.

Exp: It's ok.

Subj: Ok. So I want this. What do I do?

Exp: Press enter. Do you want that one with cell death?

Subj: No.

Exp: Ok, so go back up, and then press return again. Now you can go up to search the one

that you marked. So what do you think of these terms in relation to the term that you selected.

Subj: They seem to be pretty much related. So can you get other things besides the terms? Can you go and search?

Exp: No.

Subj: Oh, you can't. It's just to give you an idea of things that could be related? Or what?

Exp: Right, eventually, this will be linked to the actual document, but right now it is just to search terms.

Subj: So you could go and search for anything that you'd like to? So for example, I searched for this. I could go back to 'proliferation' and search for that? Oh, all right. Ok.

Exp: Keep talking while you are thinking. Speak out loud while you think.

Subj: Ok. So I can go to use this again? So if I mark one of these, will I get the same kind of, the same terms again, or will it link it to some other stuff that is more related to that specific one?

Exp: To that one specifically, right. It will bring up new terms.

Subj: Alright, so just return this? Oh, alright. Ok.

Exp: So what do you think of these terms?

Subj: Well, most of them are somehow related. You have other terms that are really very general, like 'transcripts'. Although it could be 'transcripts' in the imaginal disc of Drosophila.

Exp: So that might be somewhat related.

Subj: Yah, right. Let's see. Can I get another, another term?

Exp: Right, now press enter. And down at the bottom, you see that it tells you that you want capital letters for authors and subjects, and upper and lower case for a gene name or a function.

Subj: What is a case sensitive?

Exp: That means that it matters what the, whether it is capital or not. So this is a gene name, correct? So that would be lower case. So that's fine. So press enter, and when you are done, press q. So when it comes up with just one, that's a synonym. Is that...

Subj: I don't know. I don't know.

Exp: Can you go down the list and tell me what you think of the different terms?

Subj: One by one? Ok, so you have some that are related and some that I have no idea what they are.

Exp: Which ones are related?

Subj: Well, most of them. You have things that are, like 'positional', that's too general. Or 'cel'.

Exp: You think that is general, too?

Subj: I think so. And then, if you could keep searching here, see 'role'. That would be a neat thing to have.

Exp: Oh, you want to know the role of...

Subj: Right.

Exp: You can search this along with the gene name. You can search more than one thing together. So if you mark this one, and another one that you might want to know the role of, then search the marked ones.

Subj: Oh, alright. Ok, so... Where do I go..?

Exp: Well if you want to see the role of something, then you can search them together.

Subj: Oh, just go up ..

Exp: Mark the other one as well and then you can find out what the role is.

Subj: Now it wasn't that one that I want. Again, this one, ok let's see.

Exp: So those are terms that have both segment and role in brackets would be related to both of those.

Subj: Oh, so if, ok, so, it isn't just the word 'segment'. I thought I had marked segment polarity...

Exp: You did, you did. It just shows the first term there. So what do you think of these terms?

Subj: Looks something like segment polarity genes as well, something like this.

Exp: So you recognize those as segment poliarity genes?

Subj: I think some of them are. I don't remember them now. To tell you exactly what... Things like this word here, 'cyclic AMP phosophodiesterase'. I don't think that has anything to do with segment polarity genes. I'm not 100

Exp: That one came up with just 'role'.

Subj: Oh, 'role'. So how does this search? Anything that you have in here that has the word role will be picked up?

Exp: It's based on co-occurence, so if two words appear in the same abstract a certain number of times, then it gets a higher score than if it co-occurs less frequently.

Subj: Ok. And this is sort of the first one, (pointing to top of list), that has the highest score?

Exp: Right, because they co-occur very frequently.

Subj: Oh, Ok. But does this search 'segment polarity' and 'role' or just 'segment' and 'role'?

Exp: 'Segment polarity'.

Subj: Ok. I don't know about that. And now. Let's see. No. So that would be in capital letters? Ok. They are related, most of them. Some of them I am not sure, because I am not that familiar them. Well, the first two terms, these were just the two terms that I typed in. Just split up. And then you have a gene that is related. Uh, 'establishment' I think is somewhat relevant if you were interested in how dorsal-ventral polarity is established. But of course, it is a general term. But I think it is somehow related to it, in this instance. If I were going to figure out, or search for something, for how the pattern is established, then I would go there, of course. But things like receptor, I think that's too general. (using cursor) 'Gene product', I think that is general, but then if you want to search for a gene product that is related to that, then it becomes important. If you want to link a gene product and a gene name, like 'Toll gene', for example.

Exp: Do you want to try that?

Sure. Oh, that's right, ok. So. Good.

Exp: What about it looks good?

Subj: I suppose this is the name of the toll gene product. So that is what you are looking for. And then, well, you get words like 'product'. Too general. You get people that work with that. 'Gene': that's general. 'Protein'. That's general. And you'll be able in the future to search for the protein that would be the gene product.. So I would go down here, and I would be able to see the protein, right? So, I think even though some of the things seem too general when you first look at them, sometimes they may help you with your search by giving you context. Especially if you don't know enough. For example, 'characterization'... If you have that word alone, it doesn't mean anything. But if you are talking about characterization of the gene, then it becomes important...if you can bring the gene name and the word 'characterization' in a particular abstract.

Exp: And you'd be able to find out what they've been doing to characterize the gene.

Subj: Right, right. That's right. So now, if I am searching here, and I select this, is it going to search 'mammalian homologue'.

Exp: No. At this point, it would just search it by itself, unless you were to link it by itself.

Subj: Oh, ok. Let's see. (subject marks terms) I searched 'toll gene' and 'mammalian homologue'.

Exp: And what do you see there?

Subj: Well, people that work on them, and let's see. Some words like 'extensive', for example. It's too general maybe, but maybe it means extensive homology. So you know, it can help you with the search. And 'member, it is general, but of course, I know that probably means that it is a member of a gene family, or a family of proteins. So if you don't know anything about it, well, I guess if you don't know anything about biology, then it becomes irrelevant. But if you know something, and say you want to know more about that family of genes, or other members of the gene family, then I would go to that.

Exp: Number 14 is 'superfamily'.

Subj: 'Superfamily'. So that probably means that there are many proteins that share the same characteristics. And then with 'member', it means that this particular gene is a member of the superfamily. So I guess that evn though they are very general terms, they can help you. Maybe you didn't know that it was part of a family of genes. Now you know they probably are. So you can make some deductions, yah.

Exp: So 'toll gene' didn't come up.

Subj: No, it didn't come up. Oh, what did I do?

Exp: I don't know. Just try again. Ok. There you go. Sometimes you have to wait just a moment.

Subj: There are some names of people that work on these.

Exp: Do you recognize any of them?

Subj: No. But I don't really know. So you have functions somewhere here?

Exp: Yah, it looks like it is mostly genes and people right here. You can search down. See it says there are 62 terms.

Oh, so there are more. So how?

Exp: Just keep pushing down. Right.

Subj: So I don't know anything about this. And I want to know the function. I don't know how I would go about this. I didn't get the word function here....There is nothing here that... genes and stuff like that that we've searched before with regard to the function. But even though it is very general, sometimes you want to know what function the gene has, or the role.

Exp: So these term, like I said all come from papers, or from abstracts, and so they're dependent...the linkages are dependent on what's in those. So if nobody did characterization, then it may not appear.

Subj: It may not be there. Ok. Now up? So how would I go about searching 'defective' with Nurse Cells. That's what I'm searching. If I dont' get Nurse Cells again.

Exp: Right. Nurse Cells didn't come up again. Uh. You would have to type it in again.

Subj: I selected defective before.

Exp: Right, you can search on that one.

Subj: Or should I search again? defective and Nurse cells.

Exp: You should search again. So now it says match term 1, and when it comes up match term 2, you can put in the second term, and it will search them together.

Subj: Ok. So I can type. So term 1 is the one that I just selected? Or no?

Exp: No. You're starting from scratch now.

Subj: What do I do?

Exp: Oh, that needs to be upper case, capitals.

Subj: Can I search for 'defective Nurse cells'?

Exp: I think you should search them separate.

This one, 'cel' is too general. (Subject runs finger down list, and stops at 'female sterile'.)

Exp: Is 'female sterile' related?

Subj: Yah, the Nurse cells are from the oocyte, so it is very important for that. So if they are not normal...they feed the oocyte...so yah. So I guess if you don't have Nurse cells, then you're in big trouble. That, for example, 'actin' would also be important in the feeding step, because, there are like connections between these cells and the oocyte. And this actin sort of shapes it, the connections, and so the connections go from the oocyte to the other. So all these things seem to be related.

Exp: So then for the last step. I have two questions for you. Do you have any comments, observations, or suggestions regarding the quality of the fly thesaurus?

Subj: I don't really think I know enough to tell you things to do. I think it gives you an idea of what things can be related to the word you are searching. So for me, that I don't know anything, it would be useful.

Exp: Do you have any comments, observations, or suggestions regarding the user interface, the computer.

Subj: Yah, it's easy. And if it is easy for me, it is easy for everybody, because I have trouble with computers.

Exp: Do you have any suggestions?

Subj: I don't think so. It seems to be easy.

Exp: Ok. Thank you for your participation.

Subj: You're welcome.

Fly-Worm Traversal Experiment

Subject 2 – Worm Domain Subj: So the first one is lin-12 and Notch. So lin-12 is.. They are both developmentally important genes. And lin-12 is a C. elegans gene. And Notch is a Drosophila gene. They are important – both of them – in cell-cell interactions – so I guess that should be in the middle – during development. lin-12 is a gene that is important in gonadal development and vulval development in C. elegans. Notch is important in epidermal and neural development. They both are sort of important in making a decision of what a cell is going to become. For example, Notch is important in whether a cell is going to become an epidermal cell, or a neuron, a neural cell. The same is for lin-12, in a different tissue in C. elegans. But they are homologues, I guess. So they have related functions. And they are both members of a gene family. That means the proteins encoded by these two genes are similar. They have some important features that they share. They have EGF domains; those are extracellular domains that are important for the function, so to communicate with neighbor cells. They both do pretty much the same thing and they look pretty much the same. They have, of course, transmembrane domains, because they have to be at the membrane, both of them, to be able to function. Let's see, what else? Do you want some other things, like genes that are related to this? Ok, so for example, lin-12 is closely related to another C. elegans gene that is called glp-1, which is also a member of the same gene family we were talking about. And this is important in germ line development and it is also important in early embryonic development. And it actually can have the functions of lin-12. So if you don't have lin-12, sometimes glp-1 can function during gonadal development. And Notch is related to Delta. I don't know much about Delta, but I know it is the protein that receives the signal from Notch in the other cell. I don't remember which protein receives the signal from lin-12. Let's see, what else?

I guess it's just... I don't know. There are several families of transmembrane receptors, and I guess sometimes they have the name of the first gene that was described, the first protein that was described. So probably, when you talk about this gene family, you say the Notch-lin-12 gene family, or something like that. But I'm not sure in this case. Ok. Second pair. let -which is

lethal- 23 and sevenless gene. These genes are both important in these organisms - in C. elegans and Drosophila - and they are determinants of cell fate. What are you going to become? They are both members... steps in signal transduction pathways. That is, you get a signal from the outside of the cell, it is transmitted to the inside, and you become something... what you are supposed to. And let's see. I think, yah, they are both members of another gene family. They are receptor protein kinases. I have the name of that one. So these are also both in the cell surface, because they have to receive the signal. So that is a good place to be. Of course since they are only one step of the pathway, the one in the surface of the cell, they have to send the signals to downstream genes. And both do that, and they both do that through – I should have this in the middle, but I'm going to have to put it on both sides – through ras genes. Those are very important genes. And in C. elegans, the ras gene that receives the signal from let-23 is let-60. And I don't know the name of the Drosophila ras, but I'm sure they have a name. There are other members... other proteins that are related to sevenless, for example, boss (bride of sevenless). That is actually in another cell. So what is the C. elegans gene important for? What cell fates does the let-23 gene determine? That is primary cell fate in vulva development in C. elegans. And a similar pathway, which is the sevenless pathway, is used for the R7 cell determination in the Drosophila eye development. You see, they are going to make completely different cells, but the mechanisms are similar, so I guess you can say different tissues, similar mechanisms. In all signal transduction pathways, you have a signal that is transmitted from neighbor cells, neighbor tissues. So you have a signal. You have a receptor, usually in the surface of the cell. Sometimes it is in the cytoplasm. For example, but this is for a different pathway, but it is a signal transduction pathway for hormones. They go directly inside the cell. They are tiny molecules, and they can go through the lipid bilayer. The receptors are actually inside the cell. So they go inside, they bind, and then those receptors usually go inside the nucleus and activate gene transcription. So these have one more step. They have to have a receptor in the surface of the cell, and then there are usually a number of steps. So you have to pass on the signal to different players, to other signalling molecules inside the cell. And one of these players – it is a very important one – it is in almost every single pathway of this type, is a ras gene, a ras protein

in this case, which is encoded by a ras gene. This transduces the signal to a number of other proteins. Sometimes, in a pathway like this, it has some proteins that are shared, some elements that are shared with those, and some elements that are unique. So you have different proteins. Sometimes you can finally activate, differentially activate gene transcription. So a certain cell is going to now express the genes that are going to make that cell be a vulva cell in C. elegans. This cell is going to be an R7 in the Drosophila eye.

The last one that I have is mab-5. mab-5 is a gene is the gene in C. elegans. And Antennapedia/Ultrabithorax complex in Drosophila. And these genes that are members of this family of genes are not only in C. elegans and Drosophila, but are also important in vertebrate development. And the part of development that these genes seem to be important in determining is Axial development, axial-posterior patterning. So they were first identified in Drosophila. And then the question was... people determined that they were important in determining the segmentation pattern in early Drosophila. And then, of course, the question was Do they exist in more other, in earlier, in more primitive organisms like C. elegans. And of course, C. elegans doesn't have segments. And so if they are there, are they also important in determining the anterior-posterior axis of the animal. And indeed, they found some of these genes. So they found mab-5, which is important which is important in determining posterior structures, close to the tail. And then when more of these genes were found in Drosophila, I guess some of them, like labial, they started looking for homologues in C. elegans. They did find some. ceh-11, which is the same as egl-5. So this animal has a problem; it cannot lay eggs. So the eggs hatch inside worm. And others like ceh-13 and ceh-15. So they are all important in determining a certain portion of the animal. Some the posterior part, some the... The gene that has been studied the most is mab-5, which is important in tail structures.

So these proteins in C. elegans and in Drosophila have many things in common. They all contain homeodomains, which is the part of the protein that is formed by helix-loop-helix structures. And this can bind DNA. So it has been shown that by binding DNA, it can regulate gene expression. Ok. And for most of these, they have done experiments to show that they can

regulate gene expression. Because sometimes you can have a protein that has the potential to bind DNA, but it hasn't been shown actually that the binding can activate transcription or shut off transcription. Ok, so I guess we can say on/off transcription. I don't know if any of this is going to make any sense.

I guess one other feature that these complexes have is – these genes have – is that they are arranged on the chromosome in the way they are going to be expressed in the animal. So, I guess, and amazingly that is conserved from worms to Drosophila to vertebrates. So I guess, arrangement of genes ... conserved... activation from left to right on chromosome, according where these genes are. And they are all, the Antennapedia, Ultrabithorax, labial, Deformed, in Drosophila are ordered in a certain way on the chromosomes. And this is important, labial, for anterior structures. So that is turned on first. And then the second one is turned on, which is, for example important for abdominal structures. And then the ones that are important for more posterior structures. And they are this way on the chromosome, and they are turned on this way. And the same for C. elegans. So it has to go that way. So the ones that are important for determining head structures are going to be turned on before the ones that are important for tail structures. And they are arranged that way on the chromosome, which is amazing. And that's conserved all the way to vertebrates. I don't know. I can't think of anything else. [turned tape over].

.... DNA binding domain. Then another part of the protein itself, or a protein that can bind to the region that is responsible for the activation. So if you have this domain, it will just bind DNA. But many, many proteins that have this domain are gene regulators.

When I think of a protein like this, my first association is what they do. They do something. In this case, they determine the anterior-posterior axis. That's the first thing. And then you start thinking of how they do it. Do they do it the same way? Or do they do it differently? And these actually do it similarly. But usually, and I think it is the same for the other examples, is what they do. They do something similar in an organism. Even though ... these do almost exactly the same thing, you know, anterior-posterior patterning. The others, for example cell to cell signalling, the

general mechanisms are the same, but they do it in completely different tissues. But they do the same thing. Like for example, they are members of the same gene family, and they transmit... but then one makes the Drosophila eye cell, and the other one makes the C. elegans vulval cell. So that's why I think that the first association is what do they do. Are they similar to one another or not. That's the way I think about it. And then, you think, you know, about whether they are similar proteins or not. And then whether they have similar domains, whether they share things. But the first thing is do they do something that is similar in different animals. Somethings you can actually, you know, take a gene from one organism and put it in another, and get it to do the same thing. [long pause]

Exp: When you talk about doing something in the other organism, you are talking about...

Subj: Complementing a mutation. For example, and this is not the case here. It won't work in this case, but if we think about genes that are involved in cell cycle control... For example, a gene that is involved in cell cycle control in C. elegans, you can put it in yeast, in a yeast that has been mutant for the homologue, and now you don't have a mutant anymore. It can complement, the C. elegans gene can complement the yeast gene, for example. A protein kinase, that has certain characteristics, and they know the homologue. They are similar enough to be able to perform the same function in yeast that they do in C. elegans. It is more difficult to do here, because Drosophila transformation is very difficult, and Drosophila genes are very finely regulated. They have huge regulatory regions that do not actually encode for a gene, but they have signals for its regulation. In the promotor. So it can be at the 5'-end, or sometimes at the 3'-end, sometimes in the intron. And C. elegans is much simpler in that respect. You don't need that much extra things to make a gene work when you put it in the organism.

Exp: Actually this last bit, where you talk about your pattern of thought process is very helpful. So you look at the gross level first.

Subj: Yes. That's what I usually do. But, yah, sometimes if you are doing research, sometimes it can go the other way around. You find, you have your gene, and you find something,

like a helix-loop-helix domain. YOu have no idea what your protein is doing, what your gene is doing. But you say, ok, it has thins domain. Maybe it will be able to bind DNA. And then, you know, you start relating to your other proteins that have been described before that have the same domain. And then you say, oh, well, maybe they do something similar. And then you look in your organism, and you say, ok, can they do this? or can't they. If you are working with something normal, you know if you have a pair like that, I always think of what they do first.

Fly-Worm-WCS Document Retrieval Experiment

Subject 5 – Novice

Exp: First we will just search and see how the search engine does by itself.

Subj: Actually, so there is an author who's last name starts with Spiros.

Exp: So it didn't come up with anything, so we'll invoke the thesaurus.

Subj: Here we go. This is the guy. So he sequences this gene. And I want to see which parts he has sequenced so far.

Exp: Let's add a word down here. It's working very slowly today. Something is really taking up a lot... Now we'll re-invoke the thesaurus to try to get just Notch and his sequences, and see what terms come up all together. Now actually, what is going to come up is worm documents. So it is unlikely that he, being a fly researcher is going to have published in the worm literature.

Subj: So I see facets. That's a good one.

Exp: Do you want to add facets?

Subj: Yes. Introns is a good one.

Exp: So that is the end. So out of these, how do you want to construct the search? Would you like it to be an "and"?

Subj: Yes. Notch is a neurogenic gene. And add epidermal. That's a good word. And add neurogenic. And that should be enough. [No results produced by search.]

Exp: So are there any of these other terms that you have identified that you'd like to try?

Subj: How about neurogenic and embryo or embryogenesis. [No results produced by search.]

Exp: So how else might you construct it? So you know that Notch is a gene that is expressed during embryogenesis, and it..

Subj: And it is neurogenically active. I'm not sure if it is an EGF, but you can try.

Exp: So it came up with one item.

Subj: I'm not sure. So let's try neurogenic and genes. I guess I would want to see neurogenic genes expressed during embryogenesis.

Exp: So there are four.

Subj: That looks like a good one; that one's relevant. This one doesn't really say; it just talks about the neural system. Ok this sound good.

Exp: So if you had come up with this in a real search would you follow any of these specific genes that are mentioned to see if they had any similarity.

Subj: Yes. Uh-huh. This one is neurogenically active.

Exp: So did you know about these OPA repeats before?

Subj: No. So that is useful.

Exp: Oh, good.

Subj: That sound good. This sounds interesting too. It has similarities to repeats that are found in Drosophila, which is interesting.

Exp: So 3 out of the 4, and you have several possible genes that you can look at. So do you have another one?

Subj: So how about techniques? Can we do techniques? In situ hybridization.

Exp: What about in situ hybridization?

Subj: I'd like to see what kinds of approaches they use.

Exp: Wow, so there are 101. Do you want to narrow that down some?

Subj: Maybe with, in embryos.

Exp: Ok, maybe use embryo, and we'll so a subsearch of this set. Ok. So 21. That's a more reasonable set.

Subj: Can we narrow it down more?

Exp: Is there something more specific that you'd like to use?

Subj: Ok, maybe looking at transcription, using [in situ hybridization].

Exp: Ok, so 8 items.

Subj: So this would be looking at how they use in situ hybridization to study transcription of embryonic genes. Ok, good. So what are these? Are these articles?

Exp: These are abstracts from the Worm Breeders' Gazette. At this point we don't have copyrights, so we can't put in Medline and Biosis abstracts, so these are all abstracts from two informal publications. One is Worm Breeders' Gazette, and the other is the abstracts from the Worm Meetings.

Subj: How long does it span up to now? Is it current?

Exp: Yes it is current, and let's see. I forget when everything started, but volume 13 is within the last 2 years. Yes this one is probably '93.

Subj: Oh, this is useful. It tells you what they did. They were able to define between two RNA's using that technique. Yeah, see they have different ways of doing this. So they are doing maternal RNA's, and things like that with m-RNA. [Next one] Not really.

Exp: It is interesting that this brought up hybridisation, even though it is spelled with an "s".

Subj: Ok, this is interesting, because they use PCR in in situ's, and they are talking about that technique. That is something I would think about using in the future.

Exp: Oh, and had you thought of using that combination before?

Subj: Yeah, but I've done it only once, so by reading their work, and can see how they've done it. This is good, too. They are trying to detect what kind of genes might be transcribed early on in embryogenesis.

Exp: Is that what you are looking for? Early genes?

Subj: Yes, I'm trying to find an early gene that might be transcribed before the time period that we are looking at. So this is good. I like this because they give their protocol for in situ hybridization, and they do it completely different. They remove all the membranes, which we do in a separate step, and they include it all in one step. And this is like the same. They are trying to isolate maternal RNA from early embryonic m-RNA. That good. Not really. And last one: Yeah. This is.

Exp: Ok, by coming up with your own terms (embryo and transcription), you were able to close down on what you want to do with the in situ hybridization. So let's go to the thesaurus now. And you can go through this list and see if there are any terms that... See here is transcription and embryonic. This system would have brought those terms up.

Subj: Chromosome. Biotin-labeled. That's all.

Exp: Ok. So let's go back up here and bring up those 101, and limit it using those terms that you pulled out.

Subj: So how about using Biotin in situ hybridization with chromosomes.

Exp: So let's do it this way so that it will be this and this and this and biotin and chromosomes.

Subj: Good. The technique is interesting. They do it a little differently and they use a different gene. They are looking at metaphase chromosomes. That is different, too. Yeah, so that would be relevant. This is interesting, too, how they use it in mapping. They are doing metaphase scans.

Exp: What cell cycle phase are you looking at?

Subj: Interphase. Ok and last one. Yeah, this is pretty interesting too, because they talk about the observations they made in [unclear], and what the structure might be.

Exp: Ok, so do you have another that you would like to try.

Subj: We can do chromosome structure. Oh, these are all the things.

Exp: Yes, these are all the documents in alphabetical order. If we were to go here, you could see all the genes in alphabetical order, and just go down. Say you weren't sure what all the different possible ace genes are, then it could trigger your memory there, too.

Subj: That's neat. 102. Yeah, so they're doing a lot on chromosome structure. So let's add interphase and nuclei. [No results produced by search.] Just try interphase. [No results produced by search.] So let's try meiotic. Ok so let's look at "Age related changes..." That's interesting. Oh, let's see "Genetic..." Yeah, that's a good one. I guess that would be it.

Exp: Ok, so this one was chromosome and structure and meiotic. So we'll put in chromosome and structure, and invoke the thesaurus, and see if it can bring up some other, additional terms that might help you limit this big group.

Subj: Let's try just chromosomes.

Exp: These are the terms that are under just chromosomes.

Subj: You can see whether he worked on it in C. elegans.

Exp: Oh, ok. Let's put him up here. No. [No results produced by search.]

Subj: General terms seem to bring up general terms. It's not a good idea. How about, what was this? Chromosomes and structure. How about chromatin and structure. [Invoked thesaurus]. Histone.

Exp: Histone. Ok. What do you want to know about chromatin and its structure?

Subj: I guess I want to know to what extent they have been doing their research in this area.

Get some specifics. I guess core would be another good one.

Exp: Ok. So let's try up here [at WCS search window]. Would that make sense?

Subj: Uh-huh. That's good. Ok, so there is 1. It is an interesting one.

Exp: Ok, so are you satisfied with the outcome of that search?

Subj: Uh-huh.

Exp: Ok. so is there anything else that you would like to try?

Subj: How about heterochromatin? How about heterochromatin and embryo, to see when it becomes evident, or when it starts. [No results produced by search.]

Exp: Ok. Let's try embryonic.

Subj: Oh, so that's the same paper. Good.

Exp: Ok. Well is there anything completely different that you want to try? do you have time for one more?

Subj: Yeah. What do you mean completely different?

Exp: Well, you did Notch, and you did in situ, and you did the chromatin. Is there another category that you would like to investigate?

Subj: How about fixation? It's kind of general. It's probably vague.

Exp: And how are you using that?

Subj: I guess as a technique.

Exp: And what would you expect to be fixing?

Subj: Embryos.

Exp: As opposed to fixing your gel.

Subj: Yeah. That's a lot. Ok. How about fixation using paraformaldehyde? That's very specific. Wow, so. Oh. Cool.

Exp: You like the second one?

Subj: Yeah. Yeah, that's a good one. It's showing you how they're doing it. Probably that one.

Exp: In situ?

Subj: Yeah. I think we saw that one already.

Exp: Yeah, that one came up already. And that's a good one?

Subj: Yeah. I think that's it.

Exp: Ok. So, let's go back up here and do fixation, and go to the thesaurus. Is there anything in here that you'd like to add to limit this set?

Subj: Glutaraldehyde. Let's go down [the list] a little. Yeah, microtubules. Good.

Exp: Ok. So we have fixation over here. And how do you want to construct this search? Fixation and both of these, or do you want to try them one at a time?

Subj: Let's say, fixation to microtubules using glutaraldehyde. So there's two. The second one. That's a good one. Good. Let's look at the first one, too. That's good, just to see what concentrations they used. That's useful.

Exp: Ok, so can you describe for me how you perceive this tool, and the usefulness you see for the tool: the thesaurus in particular.

Subj: I think technique-wise, it would be useful, just to see if you haven't done something similar in Drosophila, and to see what they have done in C. elegans. Different ups and downs that they have had. Different solutions that they have used, and stuff, so that you can try it in Drosophila. So I think that would be a useful tool.

Exp: So if this were available for worm or for a number of different organisms, do you think you would use it?

Subj: Yeah, I think so.

Exp: Ok. Anything else you'd like to say?

Subj: No, it's good.

Exp: Well thanks for coming down.

Subj: I hope it was useful. Like when you said embryo and embryonic and embryogenesis, I lumped all of those together, and remember at the beginning, we had to have it as embryonic, or it didn't come out. So little specifics like that helped out.

Exp: Ok. Well thanks very much.

Subj: Sure.

Appendix C

Tables and Graphs

Expertise Level	# New Searches	# Iterations for Each New Search			
		1	2	3	4
Expert	2	3	0		
Expert	3	1	2	1	
Novice	4	3	1	2	2
Novice	3	0	0	1	
Outsider	3	1	1	0	
Outsider	4	2	2	0	0
Total	19	10	6	4	2

Table C.1: Number of Iterations per New Search Used by Subjects While Browsing the Fly Thesaurus

Object	Fly Experts			Worm Experts		
	Source	Intermediate	Target	Source	Intermediate	Target
S	6	42	4	1	45	2
G	5	5	6	7	5	6
A	0	0	0	0	3	0
Total	11	47	10	8	53	8

Table C.2: Object Types for Terms at Various Traversal Positions

Subjects	1	2	3	4	5	Total
Worm	26	8	9	1%	0	39
Experts	66%	21%	10%	3%	0%	100%
Fly	26	2	3	0	5	36
Experts	72%	6%	8%	0%	3%	100%
Overall	52	10	7	1	5	75
	65%	13%	9%	1%	3%	100%

Table C.3: Number of intermediate nodes in traversal – entire phrase

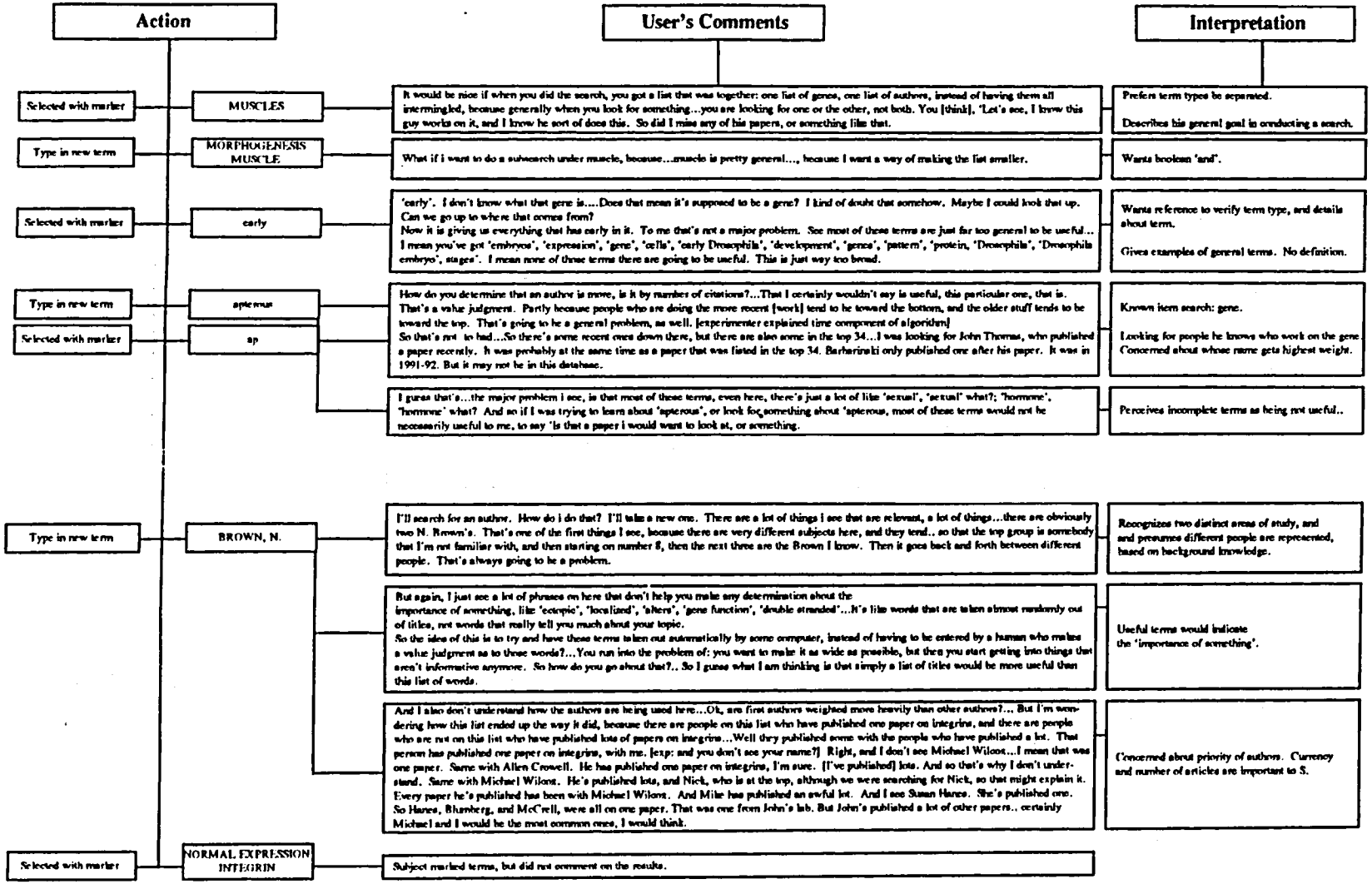
Object Type	# Found	# Not Found	Total # Queries
Gene	11 (.69)	5	16
Subject	15 (.625)	9	24
Author	3 (1.00)	0	3
Total	29 (.674)	14	43

Table C.4: Query terms found in concept space, by object type

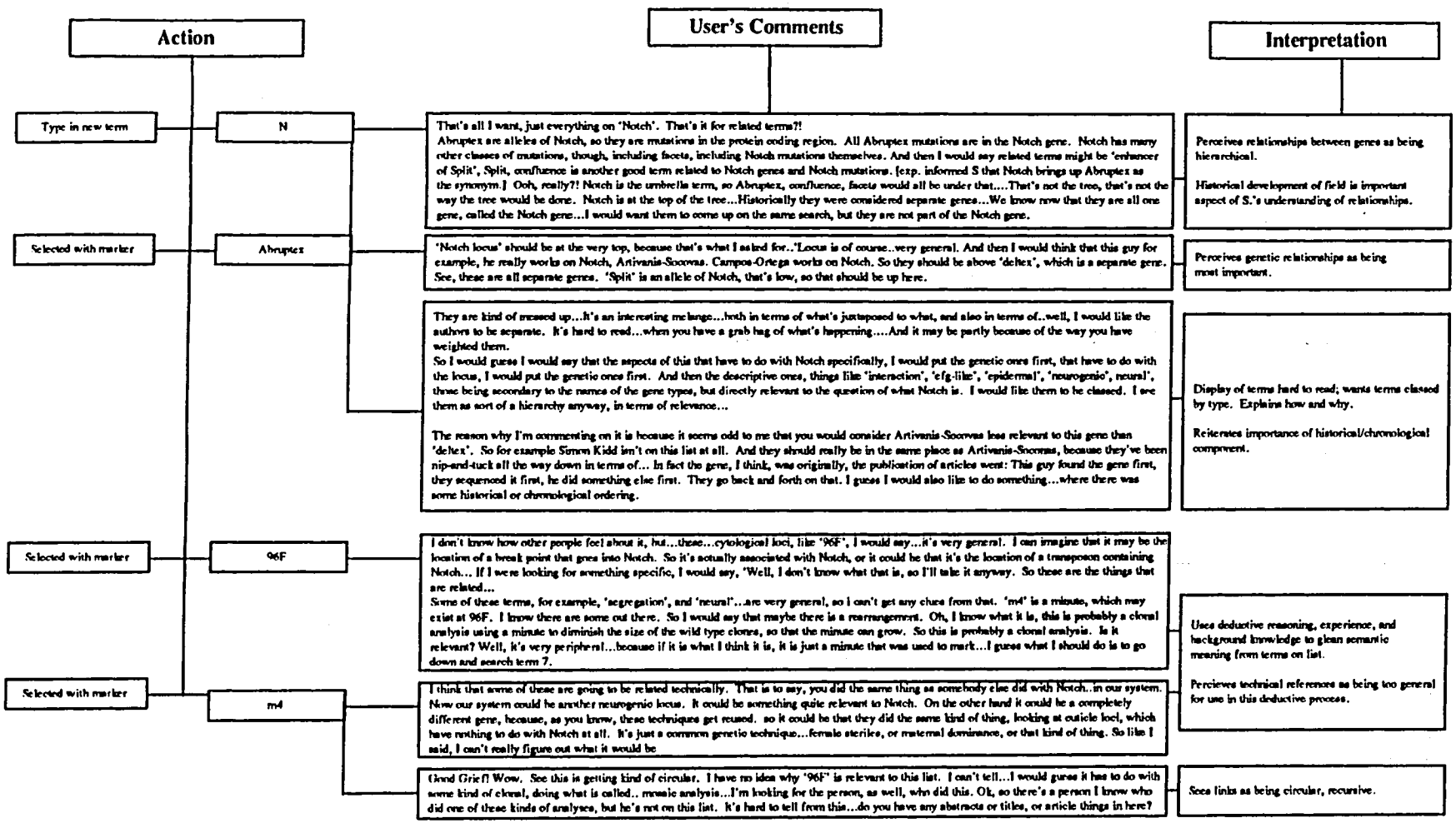
Heuristic	Number of Instances				
	0	1	2	3	4
Attempts in WCS prior to concept space activation	0	27	7	1	1
Times concept space activated per search	3	24	6	3	1
Times concept space list consulted and query reformulated per activation	18	19	8	2	3

Table C.5: Number of instances of various search heuristics using concept space

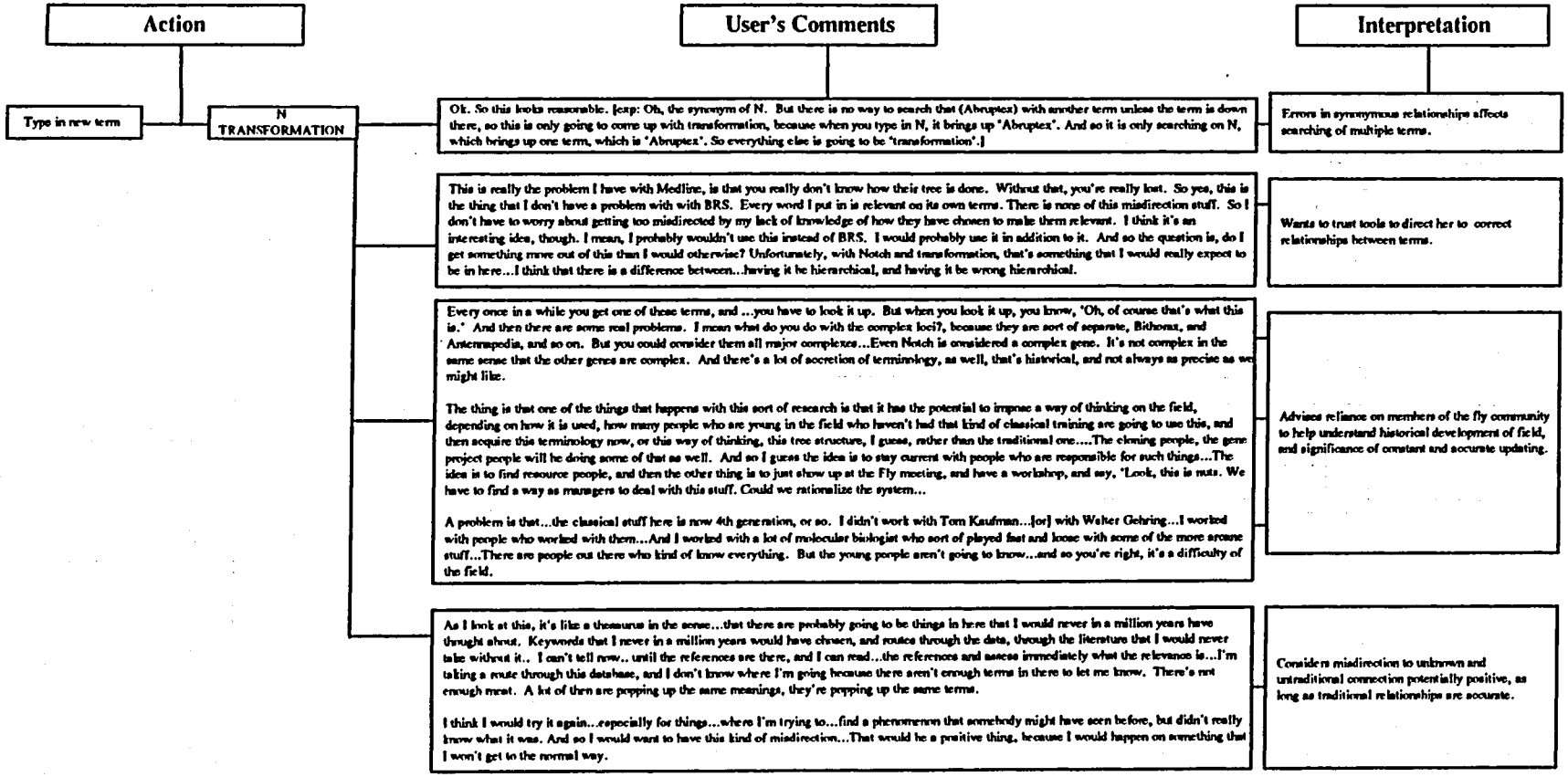
Subject 1 -- Expert



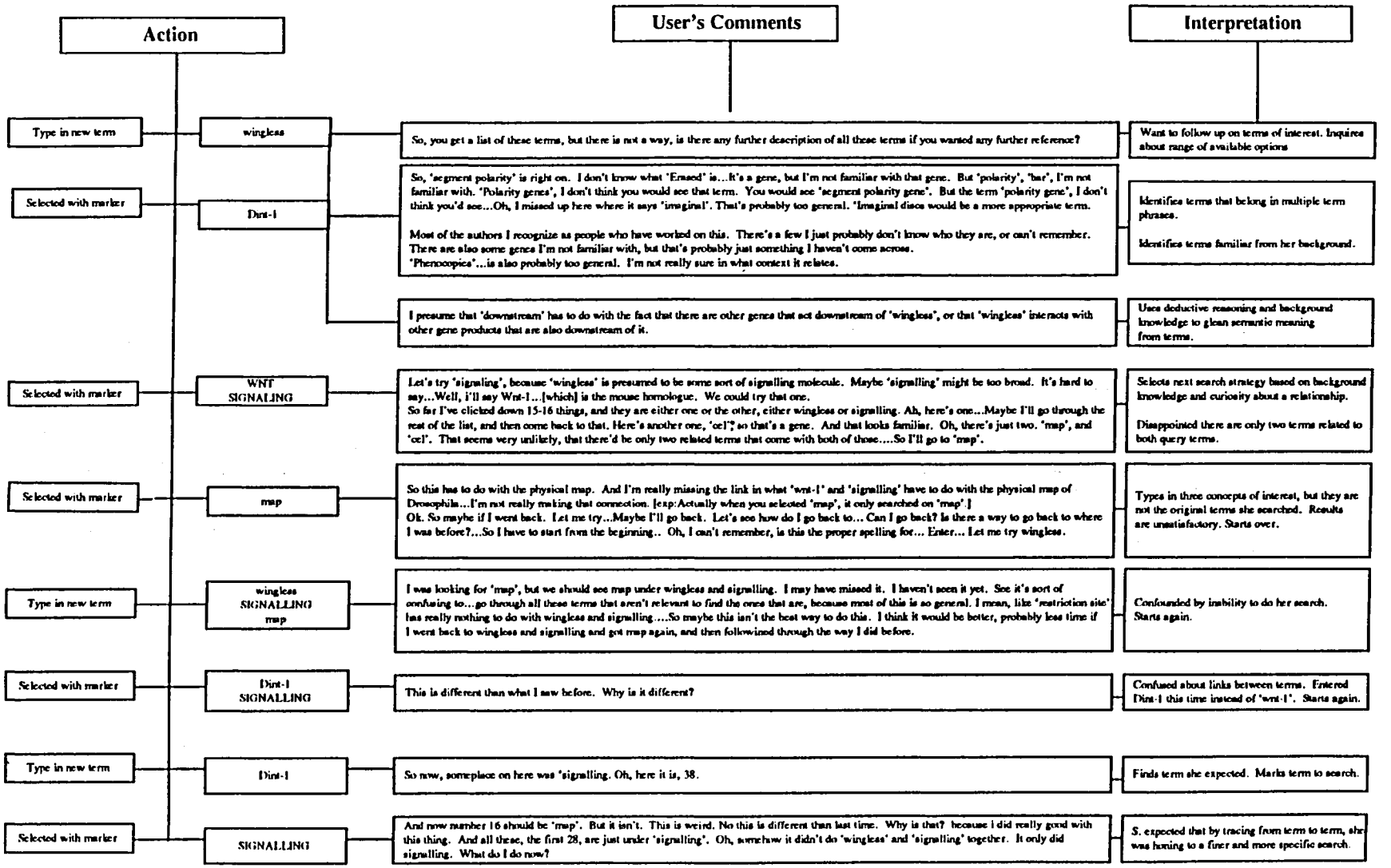
Subject 2 -- Expert



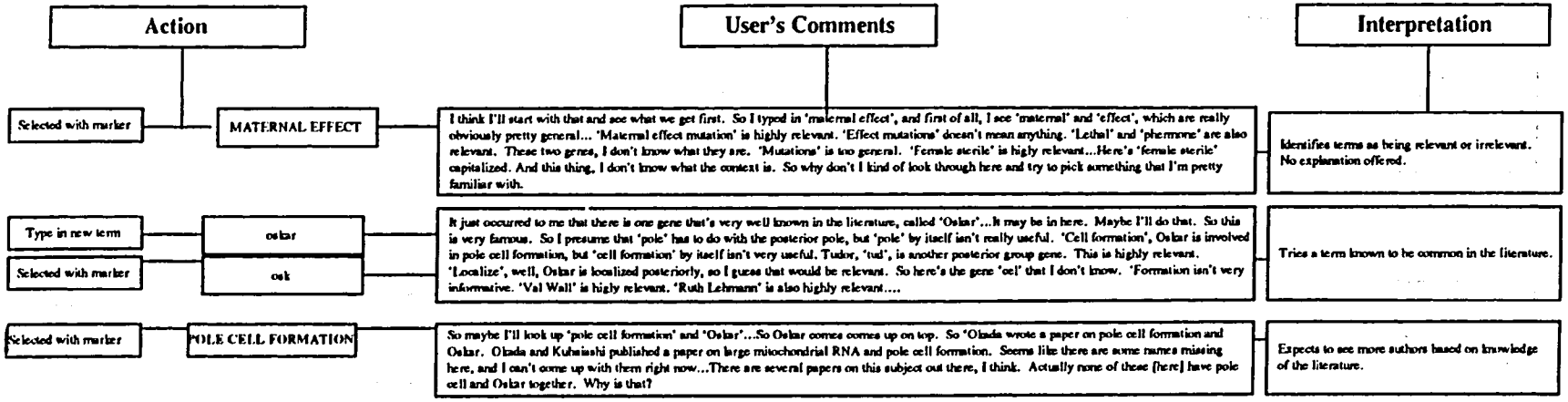
Subject 2 -- cont.



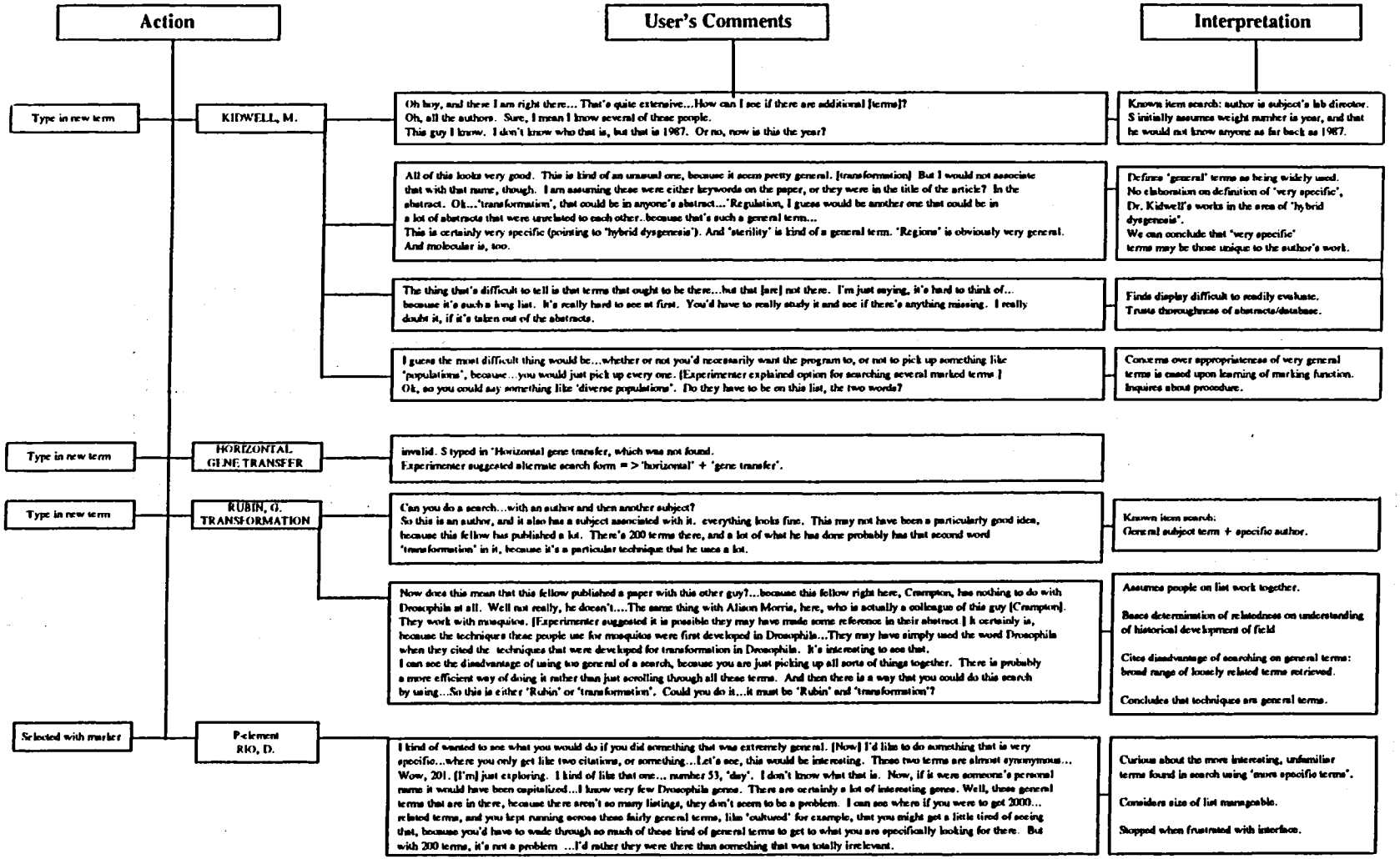
Subject 3 -- Novice



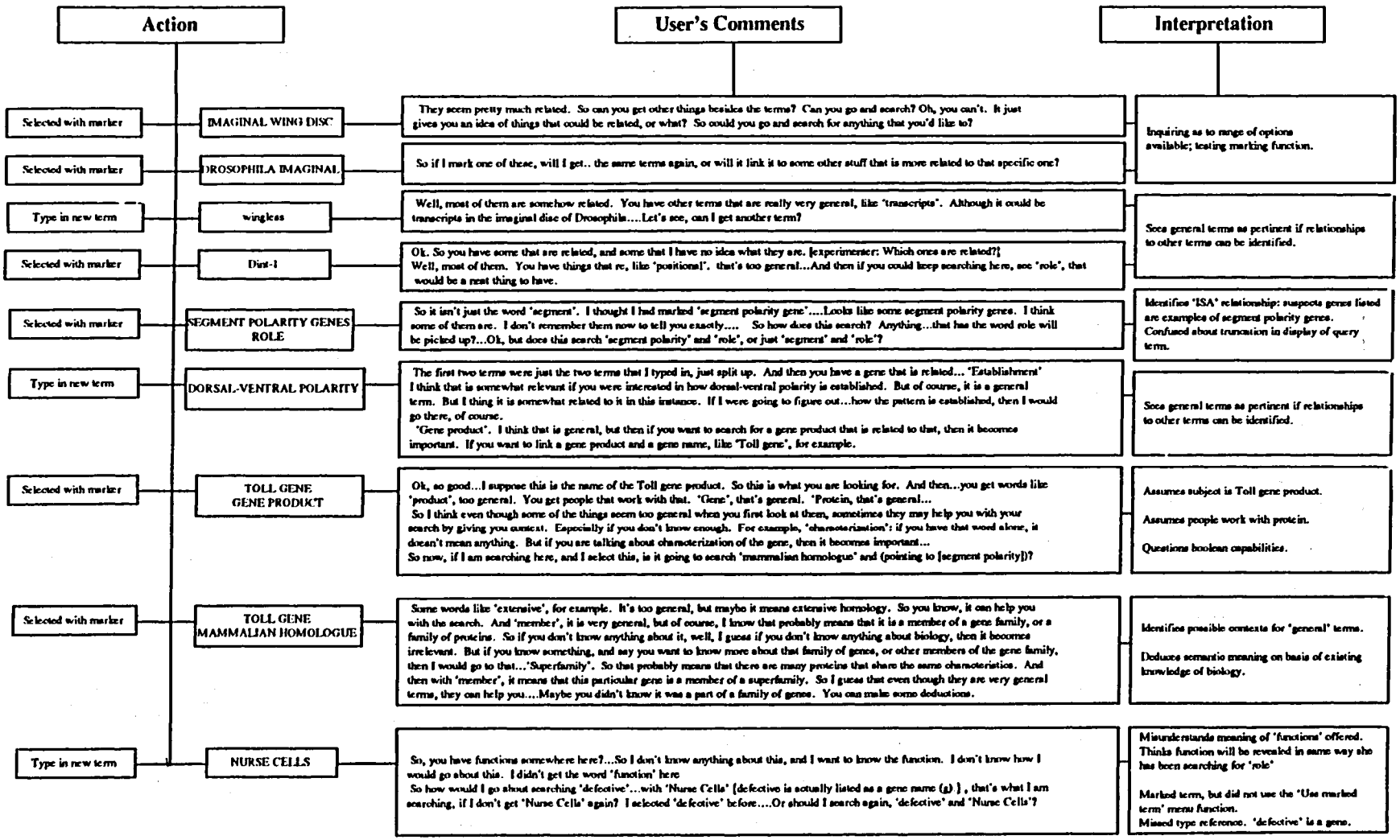
Subject 3 -- cont.



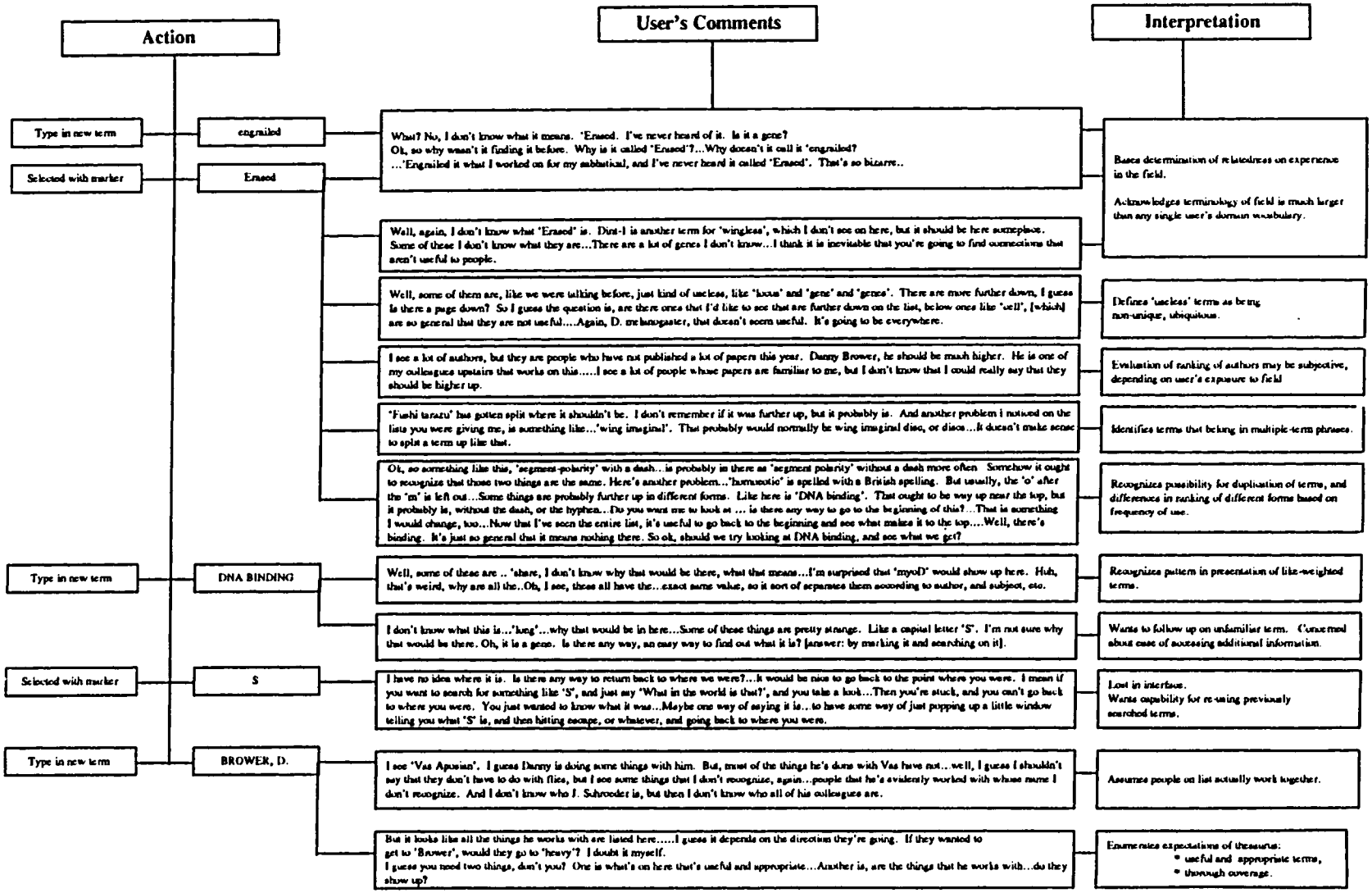
Subject 4 -- Novice



Subject 5 -- Outsider



Subject 6 -- Outsider



Appendix D

Inclusion of Manuscript for Publication

Generating a Domain-specific Thesaurus Automatically: An Experiment on FlyBase

Hsinchun Chen [1], Bruce Schatz [2], Joanne Martinez [3], Tobun Dorbin Ng [4]

February 9, 1994

[1]University of Arizona, MIS Department, Karl Eller Graduate School of Management, University of Arizona, McClelland Hall 430Z, Tucson, Arizona 85721, hchen@bpa.arizona.edu, (602) 621-4153.

[2]University of Illinois, NCSA, Beckman Institute, 405 N. Mathews Ave., Urbana, IL 61801, schatz@csl.ncsa.uiuc.edu.

[3]University of Arizona, Library Science, martinezj@bpa.arizona.edu, (602) 621-2328.

[4]University of Arizona, MIS Department, University of Arizona, tng@bpa.arizona.edu, (602) 621-2328.

Contents

1 Introduction 1

2 Scientific Databases, Electronic Community Systems, and Information Retrieval Problems 1

3 An Algorithmic Approach to Automatic Thesaurus Generation: An Overview 4

4 Generating a Fly Thesaurus Automatically 7
- 4.1 Object Filtering . . . 7
- 4.2 Automatic Indexing . . . 8
- 4.3 Cluster Analysis . . . 9

5 Current Status and Experimental Design 12
- 5.1 A Sample Session . . . 13
- 5.2 Experimental Design . . . 14

6 Results of Fly Thesaurus Evaluation 19
- 6.1 Results of the Association Experiment . . . 19
- 6.2 Results of the Browsing Experiment . . . 23
  - 6.2.1 Potential Pitfalls . . . 23
  - 6.2.2 Potential Usefulness . . . 26

7 Conclusion and Future Directions 27

8 Acknowledgments 29

List of Figures

1 Sample co-occurrence table . . . 12
2 Enter HOMEOTIC . . . 14
3 Select new terms . . . 15
4 Iterate with new terms . . . 16
5 Subject-suggested descriptors . . . 17
6 Sample system-suggested descriptors . . . 18
7 ANOVA analysis for relevant terms . . . 20
8 ANOVA analysis for concept recall . . . 21
9 ANOVA analysis for concept precision . . . 22
10 Taxonomy of system problems as identified by searchers . . . 24

Abstract

This research describes an algorithmic approach to the automatic generation of thesauri that can serve as online search aides for scientific databases or electronic community systems. Using *object filtering*, *automatic indexing*, and *cluster analysis* techniques, we extracted key terms and phrases representing Drosophila-related research knowledge from a core collection of Medline and Biosis abstracts. In addition to subject descriptors from the named databases, objects used in object filtering included gene and protein names, cellular and biological function names, researcher names, and subject descriptors from FlyBase, a database currently in use by molecular biologists in the *Drosophila melanogaster*-related research community. On average, each term had about 41 weighted (0..1) neighboring terms indicating *relevant* concepts.

We tested the fly thesaurus in an experiment with six fly biologists of varying degrees of expertise and background. This study confirms earlier research (of C. elegans worm thesaurus) that demonstrated the feasibility of an algorithmic solution to the *information overload* problem in databases and the *vocabulary* problem in online information retrieval. The knowledge and literature representing Drosophila research presents important challenges, including non-standardized naming conventions, a long history with a great deal of vocabulary fluidity, and about ten times the volume of literature representing the worm research. The experiment showed that the thesaurus was an excellent memory-jogging device and that it supported learning and serendipity browsing. Despite some occurrences of obvious noise, the system was useful in suggesting relevant concepts for the researchers' queries. The experience and lessons learned during the fly thesaurus development and evaluation process are presented in detail.

1 Introduction

Biological research is highly data-intensive, and data accumulation in this area is growing extremely rapidly because of massive efforts such as the Human Genome Initiative and other genome mapping projects. At present, the genomes of several organisms are being sequenced and mapped, including Caenorhabditis elegans (nematode worm), Drosophila melanogaster (fruit fly), Mus musculus (mouse), Homo sapiens (human), Escherichia coli (bacterium), Saccharomyces cerevisiae (yeast), and Arabidopsis thaliana (plant). Because communities in molecular biology form around organisms rather than techniques or problems, the results generated are stored in separate databases by each scientific community. Information about the identity, function, cytological and genetic location, mutations, and aberrations of any particular gene is therefore scattered among a variety of distributed, heterogeneous databases.

These systems range from elaborate, sophisticated database management systems to unstandardized machine-readable files created through text conversion. Distributed, heterogeneous databases are preferable to vast, centralized databases because of such inherent difficulties as file size, currency, updating, and retrieval. Nevertheless, the resulting problem of information scattering makes it difficult for scientists to share and compare information that exists in different files. Connectivity among databases is essential if scientists are to make full use of their contents. Adequate tools for structuring and accessing the contents of a diverse and heterogeneous collection of scientific databases are needed to unlock the knowledge embedded in them. Tools developed for this purpose must address the following problems: information overload, information scattering, vocabulary fluidity over time, unstandardized nomenclature, and semantic differences in the vocabularies used by the various subdisciplines of biology.

2 Scientific Databases, Electronic Community Systems, and Information Retrieval Problems

The Human Genome Initiative (HGI) offers tremendous challenges not only to the biology, biomedicine, and genetics research communities, but also to the information science and computer science communities. According to Courteau [7], 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. The vast amount of knowledge accumulated during the project's scientific discovery process can only be managed with the use of computing technologies that support efficient and effective storage, retrieval, and analysis of information, that foster seamless distributed scientific collaboration, and that facilitate timely information dissemination and sharing.

FlyBase [11] is a set of linked databases designed to provide the Drosophila research community with access to broad and thorough coverage of molecular, genetic, and community information. The Drosophila community is one of the oldest groups in biological research. Most of the knowledge it has generated has been recorded in two sources, the "Redbook," (i.e., "The Genome of Drosophila melanogaster," by D. L. Lindsley and G. G. Zimm, Academic Press, 1992), and the publications of the Drosophila Information Service. FlyBase gives promise of providing improved access to domain knowledge through use of computer technologies.

A consortium of Drosophila researchers, funded by grants from the U.S. National Institutes of Health (Washington) and Medical Research Council (London), guides FlyBase development and provides advice on preferred formats to groups developing allied databases and to individual contributors of data. FlyBase is accessible and searchable through Gopher client software, or obtainable via anonymous ftp from its primary archival site at the Department of Biology at Indiana University (IU). The primary database consists of a series of flat files that contain molecular and genetic data on *Drosophila melanogaster*, including the entire contents of the "Redbook."

In order to provide comprehensive access to Drosophila information, the IU Department of Biology also provides access to numerous "allied" databases developed at other institutions, each with its own focus. Included are a unified bibliography on Drosophila; pointers to nucleic acid and protein sequence databases; stock lists directing researchers to sources of clones and other biological reagents; a genetic map of Drosophila; a list of Drosophila genes sorted by function; lists of clones from European and American sources, and instructions for placing orders; a directory of Drosophila workers; and other associated databases useful to Drosophila researchers. Forums for informal communication between researchers are provided through the Drosophila Information Newsletter and a Bionet News group dedicated to Drosophila research. Each information source is resident on computers at its "home" institution. Access to the complete set of databases is organized through the IUBIO Gopher site.

In addition to various genome databases such as FlyBase, electronic community systems (ECS) have been proposed and implemented. These have drawn significant attention recently due to the rapid proliferation and advancement of computing, databases, and telecommunication technologies. An electronic community system encodes a research community's information and knowledge and provides an online environment to support the manipulation of that knowledge. An ECS enables researchers of a scientific community to enter and share community knowledge and findings in a timely manner and in a distributed environment, and thereby to function more efficiently and effectively within the community.

An advantage of this type of system over traditional databases is that an ECS enables users to browse the available knowledge easily, record their own knowledge for others to use, indicate authorizations for users to either view or annotate their own data [7], annotate entries from others' research, and form interrelationships between concepts [21]. An ECS

is much like an electronic library where users can browse for relevant information, filter out information they do not currently need, and share data that they have collected [21].

Another novel characteristic of an ECS is its ability to handle a wide variety of community knowledge, both formal and informal [21]. In order to "live effectively within a community, one must have available both formal archival material and informal transient folklore" [21]. An ECS provides both by integrating published literature about formal research findings with intermediate results, experimental protocols, laboratory notes, electronic bulletin board discussions, e-mail exchanges, and other informal data sources, and therefore is becoming an indispensable tool that allows researchers to browse, explore, and understand a vast and rapidly-changing world of scientific knowledge at the same time it creates a group memory.

The Worm Community System (WCS), which is a major NSF-funded collaboratory project, has been considered a model electronic community system [17] [22]. Constructed for *C. elegans* researchers, it 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 not only the entire community of worm biologists but also other related biology and biomedical community members [21] [7] [22]. In previous research, we developed a worm thesaurus based on the complete WCS literature [6]. The worm thesaurus was found to be an excellent memory-jogging tool and concept-based search aide and was incorporated into the WCS Release 2 made available to worm biologists in August 1993.

While the formats of the WCS and FlyBase are very different, each attempts to serve as a central, electronic "clearinghouse" for information a researcher needs to remain up-to-date. However, despite the potentially substantial benefits of being able to access, retrieve, and analyze data and information about homologues in other organisms, the use of distributed heterogeneous databases presents many significant obstacles. Besides the technical problems inherent in the use of various database formats, other sources of difficulties include information overload, information scattering, vocabulary fluidity over time, and differences in semantic meaning and nomenclature conventions between domains.

In this paper, we present an algorithmic approach to generation of a fly thesaurus. The main techniques used in our approach are presented in Section 3. Section 4 discusses in detail the algorithms and sample results for *object filtering* and *automatic indexing* of Drosophila literature and the algorithms and findings of the *cluster analysis* process. Section 5 presents the current status of our system implementation and an experiment we conducted involving subject area experts, novices, and (fly) community outsiders. Experimental results are discussed in detail in Section 6. Directions for future research are presented in Section 7.

3 An Algorithmic Approach to Automatic Thesaurus Generation: An Overview

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 domain-specific, 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 a searcher's vocabulary problems during information retrieval.

In information science, use of a thesaurus or a knowledge base for "intelligent" information retrieval has drawn significant attention in recent years. There have been many attempts to capture experts' domain knowledge for information retrieval. A few examples are described below. CoalSORT [16], 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 [12] 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 [8], consists of a group of "experts" that communicate via a common data structure called a 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 [2] incorporated a portion of the *Library of Congress Subject Headings* into the design of an intelligent retrieval system that 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 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 [15] [14]. 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 10 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 [13], or derived from 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 [6] [4] [9] [20]. Similarity coefficients are often obtained between pairs of distinct terms based on coincidences in term assignments to the documents of a collection. For example, a cosine computation can be used to generate

normalized term similarities between 0 and 1. When pairwise similarities are obtained between all term pairs, an automatic term-classification process such as a single-link or a complete link classification can group into common classes all terms with sufficiently large pairwise similarities [10] [19] [20]. The terms in the thesaurus classes can replace the initial search terms and be used to increase retrieval recall.

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

- **Object Filtering:**

  In [1], Bates 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 generally appear to be some existing lists of subject descriptors (e.g., the subject indexes at the back of textbooks), researchers' names (e.g., author indexes or researcher directories), and other domain-specific objects (e.g., genes, experimental methods, organizational names, etc.), either online or obtainable through OCR scanning. These domain-specific keywords can be used to help in automatic identification of important concepts in documents.

  In [4], we used several domain-specific controlled lists of subject keywords, researchers' names, and organizational names for indexing in a Russian computing database (with about 200 MBs and 40,000 documents). In creating the worm thesaurus [6], we utilized author indexes from literature sources, the WCS gene list, the subject index from the Worm Book, and an existing keyword list.

- **Automatic Indexing:**

  After object filtering, the texts remaining may still contain many important concepts. An algorithmic approach to the identification of remaining descriptors is required. An effective and robust method for content identification that is simple and domain-independent is the *automatic indexing* technique, often used in information science for indexing literature. In [20], Salton 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. A stop word list is then 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.

- **Cluster Analysis:**

While *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 a weight with each descriptor 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 [20]. The basic rationales underlying these two measures are that: 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* [10], 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 convert the raw data (e.g., indexes and weights) into a matrix of inter-individual *similarity*, *dissimilarity* or *distance* measures. The result of a cluster analysis will be a number of groups, clusters, types, or classes of individuals [10]. In *automatic thesaurus generation* [9] [4], 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. In the biomedical and biological domains, we have found that the subject matter is often very specific (i.e., gene names, experimental methods, and topics) and provides a sound basis for performing cluster analysis [6].

Although the above techniques had been employed in other applications, including the development of a highly domain-specific, up-to-date automatic thesaurus for the worm community, the volume of Drosophila literature and the unstandardized naming scheme utilized in the fly community offered special challenges. During our system development process, significant adaptation was required to meet the specific constraints and novel characteristics of FlyBase.

Specific concepts like gene names, function names, researchers, and subject descriptors obtained from FlyBase, and from the large number of fly-related documents in Medline and Biosis abstracts, provide the foundation for automatic thesaurus generation. FlyBase itself represents the long research history of the Drosophila community. The volume of information is estimated as 10 times that of the WCS. The automatic thesaurus generated for FlyBase uses as a knowledge source a collection of 5,854 abstracts (more than 10 MBs of textual information) from Medline and Biosis CD-ROM, with a time span between 1983 and 1993. After interviewing several fly biologists at the University of Arizona,

we chose to generate and represent the important concepts described in recent literature instead of capturing the concepts exhibited in all fly literature, because the latest concepts may be of most interest to current researchers. The experience and lessons learned during the system development and thesaurus evaluation process are reported below.

4 Generating a Fly Thesaurus Automatically

In order to generate a fly thesaurus from the fly literature, we performed object filtering, automatic indexing, and cluster analysis in order. Sample results are shown in this section.

4.1 Object Filtering

We created four lists of fly-related keywords from several directories in FlyBase and from Biosis and Medline. Unlike the unified and systematic naming convention used in the worm community, the format in Drosophila is more free-style. Capitalization is important for gene and function names. Furthermore, there is no systematic way to name alleles in Drosophila. Special characters such as left parenthesis "(", pipe "|", apostrophe "'", comma ",", slash "/", and plus sign "+", are commonly used in the naming convention. Here are some examples of those names: "(-Glycerol phosphate dehydrogenase", "Ac-SD", "Su(Ste)", "l(3)4.15", "ort1 ninaE1", "ADP/ATP translocase", "fructose-1,6-bisphosphate aldolase", "(Na+ K+) ATPase (subunit", and "Casein kinase II, | subunit". Because of this unconstrained naming convention, a parser that had been developed previously and used for several applications [4] [6] [3] was revised to handle these scientific terms. We created four groups of object filters as follows:

- **Gene names:** 14,013 unique gene names were identified from FlyBase. Some of these gene names are synonyms. With information from two synonym lists and the synonym information in FlyBase, we were able to differentiate all gene names into 5,868 unique genes or groups of gene names. Since Drosophila workers use upper and lower cases to represent dominant and recessive genes, the case sensitive characteristic was retained. For example, gene names beginning with upper case characters (e.g., "Abdominal-B") indicate dominant genes, while gene names beginning with lower case letter (e.g., "abdominal-A") indicate recessive genes.

- **Function names:** 725 function names were identified from the "function" directory of FlyBase. The function names share the same case sensitive feature with gene names. Some examples are "snRNA U1a", "snRNA U1b", "transcription factor TFIIIA-like", and "transcription factor Yp1".

- **Researchers' names:** Researchers' names were extracted from the people directory in FlyBase and from the "AU:" (author) field of each document in both Biosis and Medline. The "people" directory of FlyBase, which stores address and e-mail contacts for Drosophila workers, contained 4,039 unique researchers' names. In addition, by pre-processing all 5,854 documents in both Biosis and Medline, we were able to identify 6,674 unique author names. Any textual description in the abstract of a document that matched with these researchers' names was identified as a researcher index. This object filter was case insensitive.

- **Subject descriptors:** Subject descriptors were created from both Biosis and Medline. Subject descriptors are identified by "DE:" (descriptor) labels in Biosis and by "MESH:" labels in Medline. In total, we collected 4,996 unique descriptors. They are case insensitive.

In total, we pre-identified 23,773 terms known to be specific to the fly domains. We used these object lists to "filter" domain-specific concepts from the Biosis and Medline abstracts and index the objects to the abstracts.

identify important concepts in Biosis and Medline abstracts.

4.2 Automatic Indexing

After documents were "filtered" using the four lists, the remaining text in the abstracts was processed by the automatic indexing routine. Automatic indexing implementation was mainly based on the procedure reported in [20]. The following steps were executed in order:

- **Word identification:** This step was modified to address the peculiarities of fly nomenclature. Unlike worm terminology, which is quite "clean," fly terminology includes variable usage of case and punctuation. The word identification algorithm was altered to recognize words by monitoring spaces. Punctuation was retained. Some words were converted to upper case, depending on the type to which they belong. For gene names and function names, case was retained.

- **Stop-wording:** Next, we used a *stop word* list which included about 1,000 common function (non-semantic bearing) words such as on, in, at, this, there, etc. and *pure* verbs (words which are verbs only) e.g., calculate, articulate, teach, listen, etc. Although the list had been successfully used in several previous applications, we found that that some of these *stop words* coincided with the gene object filter defined above. For example, "with", "if", and "or" are gene name abbreviations, and so were found in both the gene object filter list and the *stop word* list. Because the *stop word* list was used to remove high-frequency words which were too general to be useful in representing document content, those coinciding terms were taken

out of the gene object filter to avoid the incorrect inclusion of many semantically unrelated terms as gene names.

- **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. Based on experience in building the worm thesaurus, we decided to form phrases which contained up to three adjacent words – our system generated 1-word, 2-word, and 3-word phrases. For example, "ADDITIVE", "GENETIC", "VARIATION", "ADDITIVE GENETIC", "GENETIC VARIATION", and "ADDITIVE GENETIC VARIATION" were generated from the three adjacent words "ADDITIVE GENETIC VARIATION" in a document. We will refer to these phrases simply as *terms* in the remainder of this article.

4.3 Cluster Analysis

After the concept descriptors for each abstract were identified, we proceeded to perform term co-occurrence analysis for all documents in the document collection. A term weighting scheme based on the Vector Space model [20] and an asymmetric similarity function [4] similar to the popular cosine function [10] 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 Biosis and Medline 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., has no descriptive power).

  Usually terms identified from the title of a document are more descriptive than terms identified from the abstract of the document. This is also the case here, as confirmed by fly experts at University of Arizona. In addition, terms identified through object filtering 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 retained automatic indexing terms that appeared more than three times, in order to remove incidental noise terms. The total numbers of unique terms generated

Type of Descriptor	Number of Terms
Genes	4,875
Functions	182
Researchers	8,349
Subject Descriptors	155,523
Total	168,929

Table 1: Number of unique terms generated from Fly literature

(through both object filtering and automatic indexing) from the fly literature are shown in the Table 1. Not surprisingly, subject descriptors constituted the largest share of the descriptors.

- **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:

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

  where N represents the total number of fly documents, 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. In a previous experiment we showed that this asymmetric similarity function represented better term association than the popular cosine function [4]. The weighting factor appearing in the equations below is a further improvement of our cluster algorithm.

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

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

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 adopted the following weighting schemes:

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

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

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).

Sample entries in the system-generated co-occurrence tables are shown in Figure 1. As shown in the co-occurrence table, "DISCS" was found to be most strongly related to researcher "BROWER, D." with a weighted probability of 0.286650. The other terms related to researcher "BROWER, D." were listed in descending order. In the second entry, gene "e(bx)" was found to be most relevant to gene "white". In the third entry, the subject "LOCALIZED" was founded to be the most relevant to researcher "LEHMANN, R". The last entry revealed that the subject "HOMEOTIC" had as its most relevant term the gene "ANTC".

- **Apply Thresholds:**

  Without setting a probabilistic threshold for the co-occurrence table, the total number of co-occurrence pairs was 811,356. Some terms may have a few thousand

1. BROWER, D. : DISCS: 0.286650
2. BROWER, D. : IMAGINAL: 0.268820
3. BROWER, D. : IMAGINAL DISCS: 0.252820
4. BROWER, D. : PS: 0.246420
5. BROWER, D. : Erased: 0.226450
6. BROWER, D. : 6-Pgd: 0.193650

⋮

1. white : e(bx): 0.099470
2. white : apr: 0.087560
3. white : y: 0.067830
4. white : LOCUS: 0.067480
5. white : TRANSPOSABLE: 0.057010
6. white : copia: 0.056630

⋮

1. LEHMANN, R. : LOCALIZED: 0.354240
2. LEHMANN, R. : osk: 0.337590
3. LEHMANN, R. : nanos: 0.337590
4. LEHMANN, R. : EMBRYONIC POLARITY: 0.337580
5. LEHMANN, R. : CELL FORMATION: 0.314820
6. LEHMANN, R. : POSTERIOR: 0.229400

⋮

1. HOMEOTIC : ANTC: 0.239970
2. HOMEOTIC : HOMEOTIC GENES: 0.170770
3. HOMEOTIC : Abd-B: 0.136610
4. HOMEOTIC : DROSOPHILA HOMEOTIC: 0.098420
5. HOMEOTIC : GENE: 0.097500
6. HOMEOTIC : EXPRESSION: 0.094800

⋮

Figure 1: Sample co-occurrence table

related concepts. The enormous volume of these data not only used a lot of memory, it might also overwhelm searchers during the thesaurus browsing process. For productive user-system interaction, only highly relevant concepts should be suggested to searchers. Based on our experience from worm thesaurus generation [6], we chose 100 as the maximum number of links for any node. The resulting fly thesaurus contained 748,253 pairs of related concepts. After applying the thresholds, the total number of unique terms found in the four sources was 18,099 (a significant portion of the less common subject descriptors was removed as a result of this process). On average, each term had about 41 relevant neighboring concepts.

5 Current Status and Experimental Design

Our prototype system, which consisted of a thesaurus generation component and a thesaurus query system, was developed in ANSI C and ran on SUN SPARK stations, DECstations, and DEC Alpha. It took 10.6 hours of CPU time to generate the thesaurus

using a DECstation 5000/120 (25 MIPS, ULTRIX-based). The resulting size of the fly thesaurus was 13.5 MBs. Users browse the thesaurus through a query interface that was previously developed for use with the worm thesaurus [6]. We present a sample fly thesaurus query session first. We then discuss an experimental design that was adopted to evaluate the usefulness of the fly thesaurus and to identify areas for improvement.

5.1 A Sample Session

A sample sequence of query sessions is shown in Figures 2, 3, and 4. After a previous session involved cell death (as shown in the main window), the user was interested in finding something about homeotic genes, a class of genes involved in body pattern formation in the developing fly. The term "HOMEOTIC" had been entered in the new query box. Although multiple terms may be searched together, the user elected to search only one term, as indicated by the lower case "q" entered for term 2 (Figure 2, bottom box). Figure 3 shows a display of relevant terms in decreasing weighted order. As expected, we see that a number of gene names have appeared on the list of relevant terms. We also see that term 2 indicates that the term "HOMEOTIC" is descriptive of certain genes. Term 24, "EMBRYONIC", confirms that "HOMEOTIC" does indeed refer to genes in the developing fly. Terms 1, 10, 11, 18, and 32 are all variations on the word "antenna". This makes sense, too. A group of genes important in the proper development and placement of the fly antennae are known to be clustered together on the chromosome, and are called "Antennapedia Complex." Furthermore, terms 3, 7, 9, 12, 14, 15, and 16 are all gene names.

The user was interested in determining whether any of the system-suggested genes are included in the Antennapedia Complex, and marked terms 1 (ANTC), 2 (HOMEOTIC GENES), and 17 (COMPLEX) for further searching. The menu choice "Use term(s) with marker" under "Thesaurus Re-Query" re-activated the thesaurus. Figure 4 shows the result of the iterative query using the three selected terms. We see that ANTC (term 1) appears to function as a member of an immunoglobin superfamily, and that the protein expressed by the gene appears to be glucose dehydrogenase. We also see that six of the gene names that appeared on the list of terms suggested for "HOMEOTIC" (3, 4, 7, 14, 15, and 16) directly relate to all three of the marked terms, as indicated in the square brackets following the co-occurrence weight. Thus we can surmise that these are all member genes of the Antennapedia Complex.

A user can iteratively query more marked system-suggested terms as many times as needed to arrive at a list of sufficiently specific terms. Alternatively, the thesaurus can be re-activated by entering new user-suggested terms. Terms selected by a user during the iterative thesaurus browsing process are recorded in a separate area. For the fly thesaurus, the abstracts from which the terms were extracted are not retrievable. The next stage of development for this thesaurus will involve traversal from the fly concept space into the previously generated worm concept space and will feature capability to retrieve worm

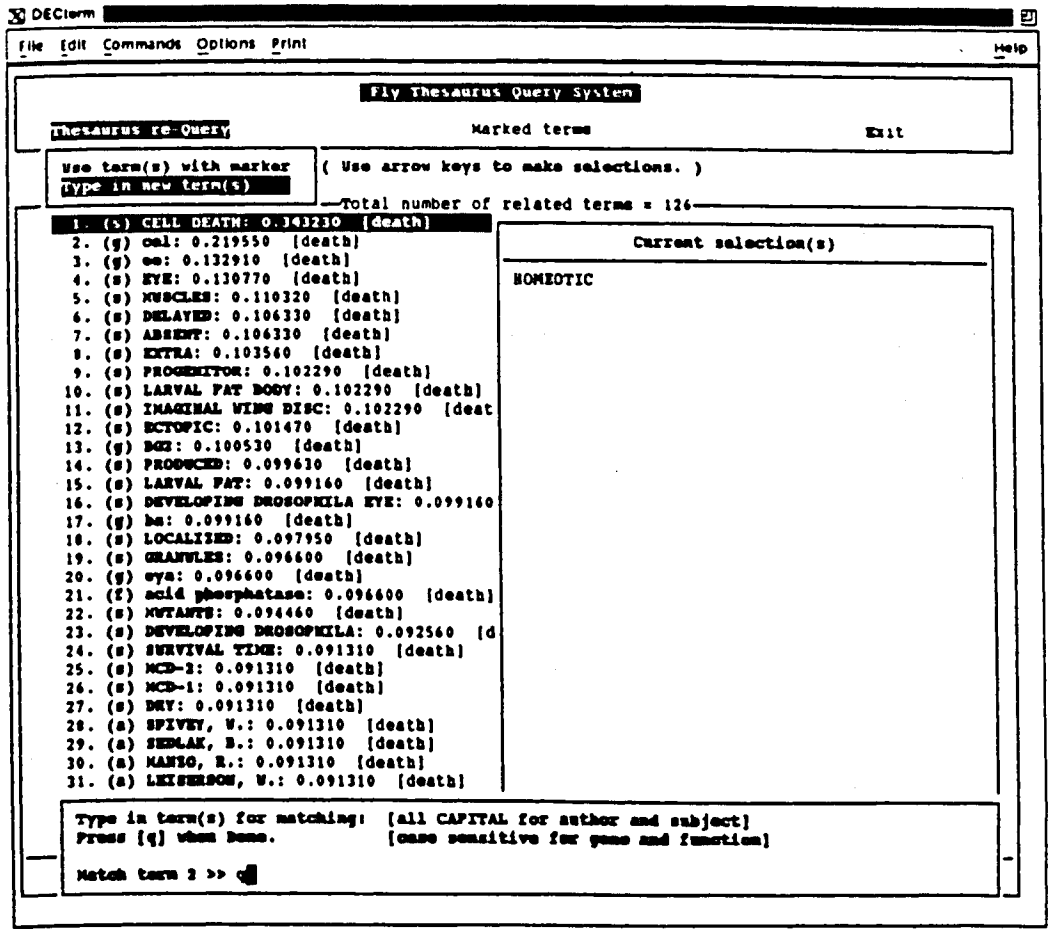


Figure 2: Enter HOMEOTIC

documents using fly query terms.

5.2 Experimental Design

A fly thesaurus evaluation experiment was conducted in Fall 1993. The experiment consisted of two parts: a term association experiment and a searcher browsing experiment. In order to investigate the effects of a searcher's background on thesaurus usage, six subjects at three levels of expertise were selected. Subjects 1 and 2 were considered experts in Drosophila biology. Both were faculty members with more than 9 years experience in Drosophila genetics, and had published numerous papers in this area. Subjects 3 and 4 were considered novices: one was a doctoral student in Molecular and Cellular Biology; the other was a postdoctoral fellow in Ecology and Evolutionary Biology. Both

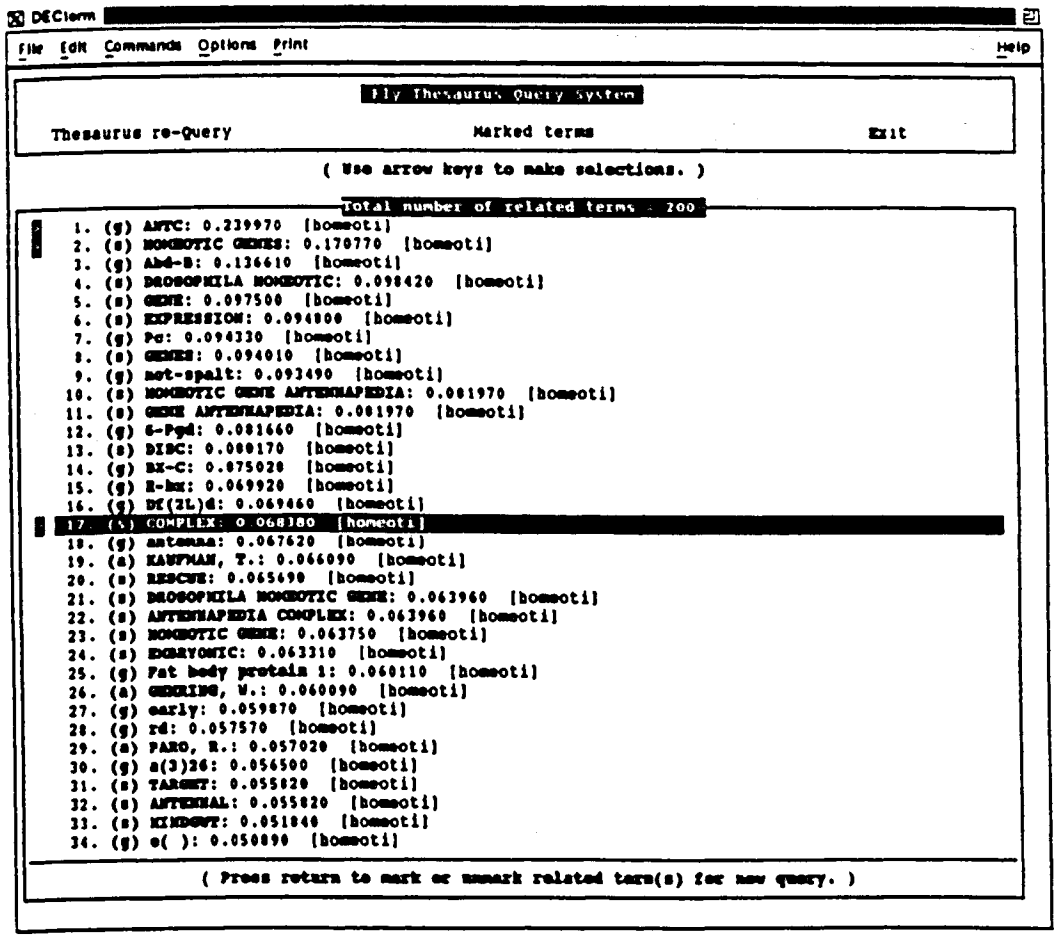


Figure 3: Select new terms

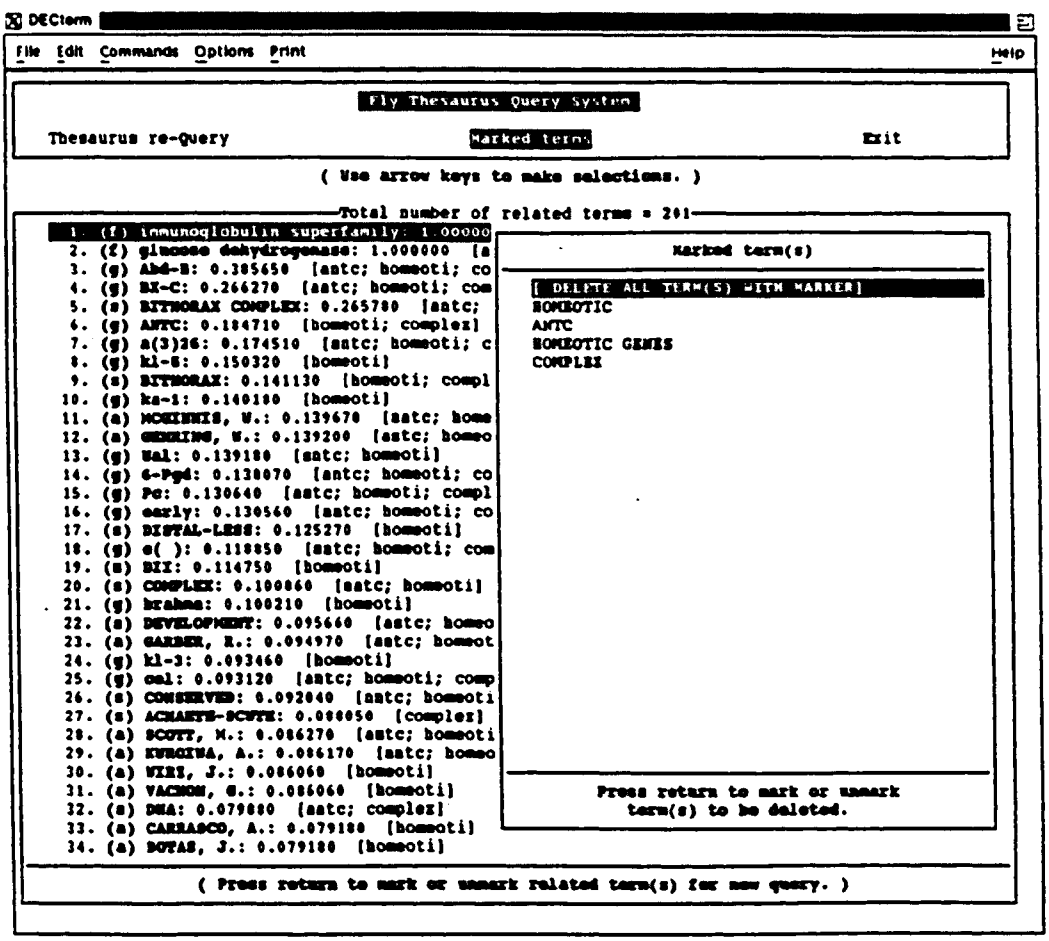


Figure 4: Iterate with new terms

Please write as many terms or concepts as possible that relate to the following term. Include Authors, Genes, Methods and Subjects that are relevant.

4) white

eyes	transvection
transport	zeste
X-chromosome	pigmentation
markers	cell autonomy
position effects	T.H. Morgan
dosage compensation	C. Bridges

Figure 5: Subject-suggested descriptors

had worked in laboratories dedicated to Drosophila research for at least 2 years. Subjects 5 and 6 were considered outsiders. Both worked outside the fly domain but had had exposure to fly concepts and had reasons to be interested in using a fly thesaurus. Subject 5 was a faculty member with limited previous experience in a Drosophila research laboratory (1 year sabbatical). Subject 6, a doctoral student with extensive experience in worm genetics, had frequently attended fly-worm joint seminars.

- **Term Association Experiment:**

  The first step of the term association experiment was to give each subject a pre-selected term. Ten fly terms chosen with the help of several fly researchers were presented to each subject in order. These terms included researchers' names, gene names, and subject descriptors. The subjects were asked to write down concepts (genes, researchers, and subject descriptors) related to each pre-selected term. A sample experimental sheet for one of the terms, "white" (a gene name)[1] and the 12 related terms generated by Subject 2 are shown in Figure 5. Subjects were then asked to mark terms suggested by the fly thesaurus as irrelevant, somewhat relevant, or very relevant. Figure 6 shows a sample system term association sheet for the same term "white" after a subject generated his own terms.

- **Searcher Browsing Experiment:**

  After the term-association experiment, subjects were asked to browse the online fly thesaurus freely – using any terms they preferred and exploring any way they liked. During browsing, subjects were asked to think aloud, and to give specific comments, observations, or suggestions regarding the user interface and the quality of the fly thesaurus. Their complete online sessions were logged. Verbal protocols were recorded and later transcribed for analysis. We aimed to identify directions for

[1] "white" is very commonly known. It was the first mutant ever found in Drosophila.

Please evaluate the term associations suggested by the Fly Thesaurus.
If not sure leave it blank. A term that is too general is considered
as irrelevant.

4) white

	Irrelevant	Somewhat Relevant	Very Relevant
e(bx)			
apr			
y			
LOCUS			
TRANSPOSABLE			
copia			
MUTATIONS			
SOMATIC			
F			
INDUCTION			
roughest	•		
brown-like			
l(2)40Fb			
Stalker			
E(wa)			
Female sterile of Konitopoulou			
UNSTABLE			
Plum			
Gart			
GENETIC			
VARIEGATION			
MELANOGASTER			
dp			
REPAIR			
ALLELES			
ELEMENT			
ELEMENTS			
M			
INSERTION			
EXCISION			

Figure 6: Sample system-suggested descriptors

system improvement through the detailed analysis of the protocols. The complete association and browsing experiment lasted 1 to 1.5 hours for each subject.

6 Results of Fly Thesaurus Evaluation

6.1 Results of the Association Experiment

- **Finding more relevant terms:**

  By counting the numbers of terms generated by the subjects themselves and the system-suggested terms marked as either somewhat relevant or very relevant by the subjects, we were able to tabulate and analyze whether the thesaurus was able to contribute relevant terms during a retrieval process. An analysis of variance procedure (ANOVA) using a statistical package MINITAB [18] 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 the fly thesaurus (KB) was able to suggest 11.200 terms. Subjects (All) were able to generate 12.017 terms by themselves. Like the findings in WCS, the expert group performed better than both novice and outsider groups. However, unlike the findings in WCS, the outsider group performed better than the novice group in generating relevant terms.

  For the fly thesaurus, the two-sample t-test revealed that there were no statistically significant differences in means (at the level of significance of 0.05) for (ALL vs. KB, P-value=0.689), (NOVICES vs. KB, P-value=0.708), and (OUTSIDERS vs. KB, P-value=0.367).

- **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. They were computed as follows:

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

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

```
                                        INDIVIDUAL 95 PCT CI'S FOR MEAN
                                        BASED ON POOLED STDEV
 LEVEL       N      MEAN      STDEV  ----------------------------------
All          60    12.017     8.288         (------------*------------)
KB           60    11.200     7.283  (-------------*------------)
                                      ----------------------------------
POOLED STDEV =     7.802              10.5       12.0       13.5

 LEVEL       N      MEAN      STDEV  ----------------------------------+-
Experts      20    17.050     7.708      (-------------*-------------)
KB           20    16.050     7.957  (-------------*-------------)
                                      ----------------------------------+-
POOLED STDEV =     7.834              15.0       17.5       20.0

 LEVEL       N      MEAN      STDEV  -+---------+---------+---------+---
Novices      20     7.200     5.167    (--------------*--------------)
KB           20     7.850     5.724       (--------------*--------------)
                                      -+---------+---------+---------+---
POOLED STDEV =     5.453           4.8        6.4        8.0        9.6

 LEVEL       N      MEAN      STDEV  ----------------------------------
Outsiders    20    11.800     8.721            (------------*------------)
KB           20     9.700     5.449   (------------*------------)
                                      ----------------------------------
POOLED STDEV =     7.272              7.5       10.0       12.5       15.0
```

Figure 7: ANOVA analysis for relevant terms

```
                                        INDIVIDUAL 95 PCT CI'S FOR MEAN
                                        BASED ON POOLED STDEV
 LEVEL       N      MEAN      STDEV    ------------------------------------
All          58    0.5381    0.2358           (-----------*-----------)
KB           58    0.4980    0.2402   (-----------*-----------)
                                        ------------------------------------
POOLED STDEV =     0.2380              0.450     0.500     0.550     0.600

 LEVEL       N      MEAN      STDEV    ------------------------------------
Experts      20    0.5576    0.2035             (------------*------------)
KB           20    0.4857    0.2081   (------------*------------)
                                        ------------------------------------
POOLED STDEV =     0.2058              0.420     0.490     0.560     0.630

 LEVEL       N      MEAN      STDEV    ------------------------------------
Novices      18    0.5286    0.2702       (--------------*--------------)
KB           18    0.5023    0.2762   (--------------*--------------)
                                        ------------------------------------
POOLED STDEV =     0.2732              0.400     0.480     0.560     0.640

 LEVEL       N      MEAN      STDEV    ------------------------------------
Outsiders    20    0.5271    0.2437       (--------------*--------------)
KB           20    0.5063    0.2477   (--------------*--------------)
                                        ------------------------------------
POOLED STDEV =     0.2457              0.420     0.490     0.560     0.630
```

Figure 8: ANOVA analysis for concept recall

For all subjects, the terms they initially generated and the terms selected from the thesaurus were included to represent the *Total Relevant Concepts* – the target set of concepts that can be obtained through user-thesaurus interaction. Based on this target set of concepts, we were able to 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, i.e., by counting the number of terms that matched the target terms. We then evaluated the *concept recall* and *concept precision* levels for the thesaurus by counting the number of thesaurus terms that 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 no significant differences (P-value = 0.366) between the subject groups and the thesaurus. On average, subjects' recall level was 53.81% while that of the thesaurus was 49.80%. These two percentages implied that the degree of overlap between the set of terms from subjects and that from thesaurus was only 3.61%. Furthermore, such findings indicated that subjects could generate almost 50% of total relevant terms independently, The thesaurus was able to help subjects associate an equal number of additional terms. Among experts, novices, and outsiders, there were no significant differences (P-value = 0.277, P-value = 0.774, and P-value = 0.791

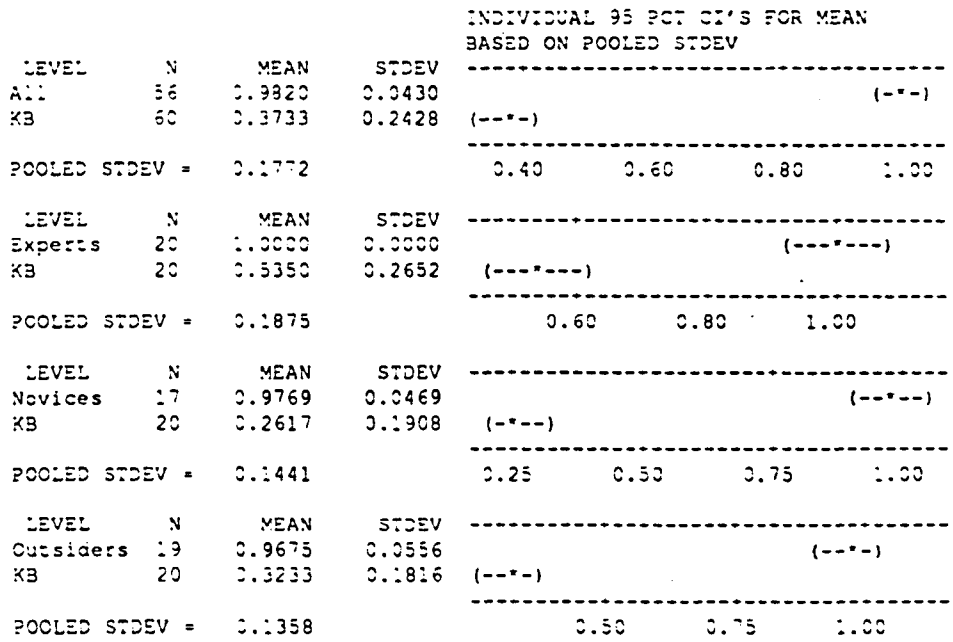


Figure 9: ANOVA analysis for concept precision

respectively) between each subject group and the thesaurus. We found that for these three subject groups, the percentage of overlapping between the set of terms generated by subjects and that generated by the thesaurus was consistent with the overall overlapping percentages. They were 4.33%, 3.09%, and 3.34% respectively. At all levels of expertise, subjects could only recall half of relevant terms without any assistance. With assistance from the thesaurus, another half of relevant terms were recalled.

As shown in Figure 9, the thesaurus produced a low level of precision compared with those produced by the human subjects. Overall, human subjects had about a 98.20% concept precision level; the thesaurus had a 37.33% precision level. The differences in overall level and all three subject-group levels were significant (P-value $\simeq 0.000$ in all levels). The low precision level of the thesaurus was attributed partially to the noise terms (mostly terms considered too general) in the thesaurus, as reported in the subjects' determination of the relevancy of those terms to their queries and needs. As is evident in information science research, even man-made thesauri are only useful when terms are 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. We believe that searchers' involvement during the thesaurus consultation process is crucial to the

success of thesaurus usage.

In conclusion, in terms of quantity, the thesaurus and the human subjects were able to generate the same number of relevant terms, but the contents of their lists were almost mutually exclusive. However, human subjects were more precise than the thesaurus. With close human-computer interaction, it appears that an automatic thesaurus-augmented search process can become very fruitful and productive.

6.2 Results of the Browsing Experiment

The subjective evaluation of the fly thesaurus reported here is the result of analyzing logged browsing sessions and transcripts of the subjects' verbal protocols. During browsing, subjects were asked to think aloud and to give specific comments, observations, or suggestions. These protocols provided clues for system improvement.

A search involved either typing in a new term (user-suggested), or marking a system-suggested term (iterations). A new search was defined as an entry of a searcher's own term. The online thesaurus suggested additional terms, which were shown on the system display. Whether they were using their own terms or system-suggested terms, most subjects used terms that they were either familiar with or curious about. Analysis of the logged search sessions revealed that the six subjects performed a total of 19 searches. All subjects performed 2, 3, or 4 new searches during their browsing sessions. The greatest number of iterations following a new search was three, with nearly 90% (17/19) of all searches being followed by 0, 1, or 2 iterations. This observation is underscored by an observation made by Subject 1, who commented that after 3 iterations she was getting no new terms/information, and that the same set of terms was reappearing. We conclude that for most searches, the system will converge after as few as 2 or 3 iterations. This has interesting implications for implementation of the spreading activation algorithms in the cross-domain traversal stage of the project.

As a result of our analysis of the session logs and verbal protocols, a taxonomy was developed to represent the observations of the subjects (see Figure 10). Included in this taxonomy are problems identified, which may be taken as potential pitfalls to consider in development of a domain-specific automatic thesaurus, and advantages/benefits identified, which may be taken as potential usefulness of such a thesaurus. This experiment gave us a better idea of how the thesaurus could be modified and used for more real-life purposes and by different user groups.

6.2.1 Potential Pitfalls

A. Quality of nodes and links:

- **Weighting of terms:**

```
''Taxonomy of User Identified Problems and Advantages in the Fly Thesaurus''

I.  Potential Pitfalls
  A.  Quality of nodes and links
    1. Weighting of terms
    2. Redundant terms
    3. Non-significant and meaningless terms
    4. Ambiguous terms
    5. Appropriateness of linkage between specific terms
    6. Database completeness
  B.  Quality of system and interface
    1. System capabilities
    2. Display of terms

II.  Potential Usefulness of the Thesaurus
  A.  Serendipity Discovery
  B.  Memory-Jogging
```

Figure 10: Taxonomy of system problems as identified by searchers

The issue of chronology was important to all subjects. However, we learned that the users had different expectations with regard to the emphasis placed on recent and seminal work, and on frequency of citation. This strongly affected their evaluation of the system-generated thesaurus. For example, two subjects (an expert and an outsider) were more concerned about the priority of more recent publications, citing numbers of publications by each person on the list, and the co-authors and recency of each. One subject (an expert) was more concerned about the priority of the authors responsible for the more seminal works.

- **Redundant terms:**

  Spelling and hyphenation of words is not standardized in the abstracts, and this created redundancies in the thesaurus. Spelling differences such as "Homeodomain" (American) and "Homoeodomain" (British), "discs" and "disks," "signaling" and "signalling" were often confusing for the user.

- **Non-significant and meaningless terms:**

  Subjects could find a term non-significant or meaningless if it is: an inappropriately included stopword; a single term from a term phrase (a fragmented phrase); or a very broad or non-specific term. The distinction between general and sufficiently specific terms was commonly emphasized by all subjects. One expert defined useless terms as those that are incomplete and those that do not indicate the "importance of something." Four subjects (two novices, one outsider, and one expert) defined specific terms as being those that would be unique in the literature, those that would clearly identify an author's field of study, or those that more

clearly identified the context of a more general term.

- **Ambiguous terms:**

  One source of ambiguity was caused by the presence of multiple authors with same name. One outsider performing an author search was perplexed by the output because the terms seemed so disparate, until he concluded that the list of terms represented the work of more than one author. Although the interface includes a field to indicate the term type (subject (s), gene (g), function (f), author (a)) to reduce ambiguity, most users did not pay much attention to it. Consequently, some terms that fit into more than one "type" were considered by the subjects to be ambiguous.

- **Appropriateness of linkage between specific terms:**

  A problem related to synonymous linkages between genes was identified. We used two of the flat files in FlyBase that linked gene names with all synonymous names. One of the subjects (an expert) pointed out that some of the genes listed as synonyms are not actually synonyms, but are alleles, facets, etc. One term (Abruptex) with weight of 1.000 (synonym) was retrieved for the Notch query. The subject pointed out that Abruptex is a class of alleles that fit under the Notch umbrella. Therefore Notch should be at the top of the tree. However, these two terms were listed in the FlyBase as synonyms. Our use of the Synonyms files from FlyBase resulted in incidents in which the thesaurus brought up gene names unfamiliar to the subjects. All 6 subjects pointed out that they were unfamiliar with "cel", a gene name that came up in at least one search for each subject, and questioned the appropriateness of this linkage.

- **Database completeness:**

  The two experts and the outsider who is a faculty member were frustrated by not seeing names they expected to be on the list. One novice was dismayed to see that the gene she works on ("cappucino") was not in the thesaurus. We estimate that because of the field's long history, the amount of formal and informal literature representing the area of Drosophila genetics is at least 10 times that for the area of worm genetics. The database underlying the thesaurus is only a sampling of the fly literature, fewer than 6000 abstracts from 2 databases.

B. Quality of System and Interface

- **System capabilities:**

  Ease of use of the interface was an issue for all of the subjects. Half of the subjects were finally comfortable with the interface by the end of the browsing session (including one who stated at the beginning of the session that she "hates

computers"). Half never became comfortable, but two of these felt that with more experience its operation would eventually become "second nature."

- **Display of terms:**

  The query term is not displayed at the top of the list of retrieved terms. This causes problems when the searcher sees a term lower on the list that is of interest and wants to search it together with the original term. To do so requires that the user re-enter both terms using the "Type In A New Term" function. Several subjects were frustrated with this limitation on further searching. Both experts would prefer that terms be classed according to "type," with all authors listed together, subjects together, genes together. One subject questioned whether it is realistic to rank an author more highly related to a term than a subject or a gene name. All but one subject (an outsider) found the display "hard to read" because of the interspersion in term "types" in the display.

6.2.2 Potential Usefulness

Verbal protocols revealed two primary, novel areas of potential usefulness.

A. Serendipity Discovery:

First is serendipity discovery. One expert initially was concerned about what she termed the "misdirectedness" of the the thesaurus' search method and expressed frustration with it. However, after a few searches, she conceded that, "It's like a ... word thesaurus in that there are probably going to be things in here that I absolutely never would have thought about... and routes though the data, through the literature that I never would take without it."

Most users thought the thesaurus would be useful in introducing them to various aspects of the domain that were beyond their present level of experience. One expert and one novice found that by imagining a scenario in which multiple terms were related resulted in deductive arrival at possible contexts for seemingly unrelated terms. We have not seen this kind of creative use of the thesaurus reported before. Two subjects (an expert and an outsider) did not see this potential and were primarily interested in seeing the things they knew to be related and relevant. Anything unexpected was viewed as noise.

B. Memory-Jogging:

The second novel usefulness for the fly thesaurus is in memory-jogging. One novice searched for the gene "wingless". The term list retrieved contained the term "signalling", which triggered a spark of recognition in the subject. She said, "Let's try 'signalling', because 'wingless' is presumed to be some sort of signalling molecule." In one case, serendipity discovery and memory-jogging occurred almost simultaneously. The subject recognized the terms "m4" and "96F" on the list of retrieved terms, and was reminded that, "m4 is a minute, which may exist at 96F; I know there are some out there [the

chromosome]. So I would say that maybe there is a rearrangement. Oh I know what it is, this is probably a clonal analysis using a minute to diminish the size of the wild type clones, so that the mutant clones can grow." By applying deductive reasoning to the thesaurus's inherent characteristics emphasizing knowledge discovery and memory-enhancement, the subject's creativity was triggered. This research has demonstrated that automatic thesauri have the potential to offer improvements to searching that exceed those possible with manual thesauri.

In summary, the results from the experiment were very encouraging. The thesaurus 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. We are also in the process of re-generating another version of the fly thesaurus by considering the comments made by the subjects, e.g., removing terms which are subsets of other terms, retaining only a singular term if plurals exist, removing hyphenation, etc. The final version will be released for internet Gopher search in Spring 1994.

7 Conclusion and Future Directions

Our most immediate future research directions, as part of 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 include the following:

- *A concept space approach to solving the vocabulary problem.*

  We believe we are moving closer to finding a solution to one of the most challenging problems in IR – 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). Now that we have created concept spaces for the fly and worm communities, adopting a multiple-thesauri consultation process (we have developed one, reported in [4]), a searcher's fly-specific terms should make it possible to traverse the two concept spaces and eventually converge towards specific terms in the (target) worm concept space. Results of some initial structural comparison between the fly and worm thesauri are summarized in Table 2.

  Before devising a mechanism for traversal from one thesaurus to the other, it was important to determine the extent of overlap between the index terms contained in each. Because the object filter lists available for the two domains differed, some direct comparisons were not possible. Table 2 shows the numbers of terms in the worm and fly thesauri. These include the number of author terms, number of gene terms, number of subject terms, number of function terms (fly thesaurus only), and number of method terms (worm thesaurus only). The last three columns report

	Worm terms	Fly terms	Overlapping terms	Percent of worm terms	Percent of fly terms
Authors	2095	7221	252	12.0	3.4
Genes	845	4875	0	—	—
Subjects	4691	5821	1503	32.0	25.8
Functions	n/a	182	n/a	—	—
Methods	12	n/a	n/a	—	—
Total	7657	18101	2203	28.8	12.2

Table 2: Structural comparison of the fly and worm thesauri

the number of terms appearing in both thesauri and the respective proportion of each thesaurus that overlapping terms represent. It is not surprising that no overlap exists in gene names: the naming conventions for the two domains are extremely different. Furthermore, it is noteworthy that 252 author names appear in both thesauri. The format for author names is last name and first initial, which could present some ambiguity. Still it is likely that some authors have published in both domains. The extent of overlap for the subject descriptors was greater than 25% for the fly thesaurus and 32% for the worm thesaurus. With this much overlap, the likelihood of finding intermediate terms for concept space traversal is promising. We are in the process of designing a concept traversal experiment with the help of several molecular biologists who are knowledgeable in both fly and worm biology.

- *Other research issues: automatic thesaurus consultation, incremental thesaurus generation, and the fluidity of concepts.*

  As an extension of the current research, we will be testing several AI-based general search algorithms (e.g., branch-and-bound and Hopfield network [5]) for automatic thesaurus consultation. We hope these algorithms will assist searchers in traversing domain-specific *concept spaces* 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.

  In the current version, the thesaurus was generated in a batch mode. Because scientific databases are rapidly and continually growing, an effective method for incremental update for the thesaurus is needed. We are currently developing an incremental version of our cluster algorithm for implementation in the worm and fly thesauri. Storing some intermediate results for term frequencies and inverse document frequencies should, we believe, make possible incremental update of a thesaurus.

  So far we have not included in our analysis the "time" dimension of the documents and concepts. By time-tagging each concept and weighting concepts differently in

the thesaurus generation process, we believe a more fluid and time-precise thesaurus can be created.

The rationale behind our research is that instead of letting knowledgeable information specialists (knowledgeable in several subject areas) perform term matching and consultation for different users, we could automatically create different domain-specific thesauri tailored to the vocabularies and concepts exhibited in the related disciplines and develop an online search aid to bridge the vocabulary differences. We hope, by expanding the concept (terminology and linkage) coverage of the underlying databases, information retrieval systems will eventually be able to assist in seamless and "intelligent" concept-based information retrieval.

8 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 NSF CISE Special Initiative on Coordination Theory and Collaboration Technology, IRI-9015407, 1990-1993 (B. Schatz et al., "Building a National Collaboratory Testbed"). We would also like to thank the faculty and students of the Molecular and Cellular Biology Department, Ecology and Evolutionary Biology Department, Anatomy Department, and Biochemistry Department at the University of Arizona for their kind assistance and valuable suggestions, in particular, those of Dr. Samuel Ward, Dr. Danny Brower, Dr. John Clark, Dr. John Little, Dr. Lynn Manceau, Dr. Mary Rykowski, Ms. Alicia Minniti, and Ms. Lisa Werner.

References

[1] M. J. Bates. Subject access in online catalogs: a design model. *Journal of the American Society for Information Science*, 37(6):357–376, November 1986.

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

[3] H. Chen, P. Hsu, R. Orwig, L. Hoopes, and J. F. Nunamaker. Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 1994, in press.

[4] H. Chen and K. J. Lynch. Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5):885–902, September/October 1992.

[5] H. Chen, K. J. Lynch, K. Basu, and T. Ng. 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, April 1993.

[6] H. Chen, B. Schatz, T. Yim, and D. Fye. Automatic thesaurus generation for a scientific community system. In *Journal of the American Society for Information Science*, 1994, forthcoming.

[7] J. Courteau. Genome databases. *Science*, 254:201–207, October 11, 1991.

[8] W. B. Croft and R. H. Thompson. I^3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6):389–404, 1987.

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

[10] B. Everitt. *Cluster Analysis*. Second Edition, Heinemann Educational Books, London, England, 1980.

[11] FlyBase. *The Drosophila Genetic Database*. Available from the ftp.bio.indiana.edu, network server and Gopher site, 1993.

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

[13] F. Hayes-Roth, D. A. Waterman, and D. Lenat. *Building Expert Systems*. Addison-Wesley, Reading, MA, 1983.

[14] D. A. Lindberg and B. L. Humphreys. 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, November, 4-7 1990.

[15] A. T. McCray and W. T. Hole. 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, November, 4-7 1990.

[16] I. Monarch and J. G. Carbonell. CoalSORT: A knowledge-based interface. *IEEE EXPERT*, pages 39–53, Spring 1987.

[17] R. Pool. Beyond database and e-mail. *Science*, 261:841–843, 13 August 1993.

[18] B. F. Ryan, B. L. Joiner, and T. A. Ryan. *MINITAB Handbook, 2nd Edition.* PWS-KENT Publishing Company, Boston, MA, 1985.

[19] G. Salton. Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4):321–346, December 1978.

[20] G. Salton. *Automatic Text Processing*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

[21] B. Schatz. Building an electronic community system. *Journal of Management Information Systems*, 8(3):87–107, Winter 1991/1992.

[22] B. Schatz. Building collaboratories for molecular biology. In *National Collaboratories: Applying Information Technology for Scientific Research*, National Research Council, National Academy Press, Washington, DC, 1993.

References

Ahlswede, T. and 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.* W. H. Freeman and Company, New York, NY.

Anderson, J. R. (1985b). Indexing systems: extensions of the mind's organizing power. *Information and Behavior,*, 1.

Bates, M. J. (1986). Subject access in online catalogs: a design model. *Journal of the American Society for Information Science*, 37(6):357–376.

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

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

Burton, R. R. (1976). *Semantic grammar: A technique for efficient language understanding in limited domains.* doctoral dissertation, Computer Science Department, University of California, Irvine, CA.

Card, S. K., Moran, T. P., and Newell, A. (1983). *The Psychology of Human Computer Interaction.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Carmel, E., Crawford, S., and Chen, H. (1992). Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5):865–884.

Chen, H. and 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)*, pages 285–289, Seattle, WA.

Chen, H. and 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., and Nunamaker, J. F. (1994a). Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 37(10):56–73.

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

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

Chen, H., Schatz, B. R., Martinez, J., and 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., and Fye, D. (1995, in press). Automatic thesaurus generation for an electronic community system. In *Journal of the American Society for Information Science*.

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

Croft, W. B. and Das, R. (1990). Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the 13th Conference on Research and Development in Information Retrieval*, pages 349–365.

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

Crouch, C. J. and 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*, pages 77–88, Copenhagen, Denmark.

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

Daniels, P. J. (1986). *The User Modelling Function of an Intelligent Interface for Document Retrieval Systems.* In B.C. Brookes (Ed.), Intelligent Information Systems for the Information Society, Elsevier Science Publishers B.V., North-Holland, Amsterdam.

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

Ekmekcioglu, F. C., Robertson, A. M., and 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.* Second Edition, Heinemann Educational Books, London, England.

Feigenbaum, E. A. (1977). The art of artificial intelligence: themes and case studies knowledge engineering. In *International Joint Conference of Artificial Intelligence,* pages 1014–1029.

Fillmore, C. J. (1968). The case for case. In *Universals in Linguistic Theory,* pages 1–88, ed. E. Bach and R. T. Harms, Holt, Rinehart, and Winston, New York.

Foss, C. L. (1989). Tools for reading and browsing hypertext. *Information Processing and Management,* 25(4):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(4):341–366.

Fox, E. A., Nutter, J. T., Ahlswede, T., Evens, M., and Markowitz, J. (1988). Building a large thesaurus for information retrieval. In *2nd Conference on Applied Natural Language Processing, Association for Computational Linguistics,* Pages 101-108, Ballard, Bruce, Editor; 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., Landauer, T. K., Gomez, L. M., and 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., and 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.

Hancock-Beaulieu, M. (1992). Query expansion: advances in research in online catalogues. *Journal of Information Science,* 18:99–103.

Harman, D. (1988). Towards interactive query expansion. In *Proceedings of the Eleventh Annual International ACMSIGIR Conference on Research and Development in Information Retrieval,* pages 321–331, Grenoble, France.

Hayes-Roth, F., Waterman, D. A., and Lenat, D. (1983). *Building Expert Systems.* Addison-Wesley, Reading, MA.

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.

Ide, E. and Salton, G. (1971). Interactive search strategies and dynamic file organization in information retrieval. In *The Smart Retrieval System – Experiments in Automatic Document Processing*, pages 373–393, G. Salton, Editor, Prentice-Hall Inc., Englewood Cliffs, NJ.

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

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

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

Lancaster, F. W. (1986). *Vocabulary Control for Information Retrieval.* Information Resources Press, Arlington, VA.

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

Lindberg, D. A. and 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*, pages 121–125, Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

McCall, F. M. and Willett, P. (1986). Criteria for the selection of search strategies in best-match document-retrieval systems. *International Journal of Man-Machine Studies*, 25:317–326.

McCray, A. T. and 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*, pages 126–130, Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Minker, J., Wilson, G. A., and 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.

Newell, A. and Simon, H. A. (1972). *Human Problem Solving.* Prentice-Hall, Englewood Cliffs, NJ.

Palmquist, R. A. and Balakrishnan, B. (1988). Using a continuous word association test to enhance a user's description of an information need: a quasi-experimental study. In *Proceedings of the American Society for Information Science (ASIS) 51st Annual Meeting*, Pages 160-163, Borgman, Christine L. and Pai, Edward Y.H. eds.; Medford, NJ: Learned Information, Inc.

Peat, H. J. and 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.

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

Ramaprasad, A. (1987). Cognitive process as a basis for MIS and DSS design. *Management Science*, 33(2):139–148.

Rasmussen, E. (1992). Clustering algorithms. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Editors, Prentice Hall, Engelwood Cliffs, NJ.

Rich, E. (1983). Users are individuals: Individualizing user models. *International Journal of Man-Machine Studies*, 18(3):199–214.

Rijsbergen, C. J. v., Harper, D. J., and Porter, M. F. (1981). The selection of good search terms. *Information Processing and Management*, 17:77–91.

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

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

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

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley Publishing Company, Inc., Reading, MA.

Salton, G. and Lesk, M. E. (1971). Information analysis and dictionary construction. In *The Smart Retrieval System – Experiments in Automatic Document Processing*, G. Salton, Editor, Prentice-Hall Inc., Englewood Cliffs, NJ, 115-142.

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

Shoval, P. (1981). *An Expert Consultation System for a Retrieval Data-Base with a Semantic-Network of Concepts*. U. of Pittsburgh, Ph.D. thesis.

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

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

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

Woods, W. A. (1972). An experimental parsing system for transition network grammars. In *Natural Language Processing*, pages 113–154, ed. R. Rustin, Algorithmics Press, New York, NY.