Chapter 7

CHARACTERIZING BIOMEDICAL CONCEPT RELATIONSHIPS

Concept Relationships as a Pathway for Knowledge Creation and Discovery

Debra Revere and Sherrilynne S. Fuller

Department of Medical Education and Biomedical Informatics,University of Washington, Seattle, WA 98195

Chapter Overview

The importance of biomedical concept relationships and document concept interrelationships are discussed and some of the ways in which concept relationships have been used in information search and retrieval are reviewed. We look at examples of innovative approaches utilizing biomedical concept identification and relationships for improved document and information retrieval and analysis that support knowledge creation and management.

Keywords

concept relationships; document representation; information management; information search and retrieval; indexing systems; textual analysis

The process of tying two items together is the important thing.

Vannevar Bush, *As We May Think*, 1945

1. INTRODUCTION

Advances in the biomedical sciences have been accompanied by an overwhelming increase in the biomedical literature. It has become critically important to not only understand developments in one's own area of specialization, but to also be able to learn quickly about developments in related and occasionally unrelated subject areas. The ability to rapidly survey the literature and integrate information gathered by researchers from multiple fields of expertise constitutes the necessary first step toward enabling biomedical scientists and researchers to keep current in their field.

Interest in developing techniques and methods for processing documents and document collections, with the goal of providing the information most relevant to a user's need, pre-dates widespread use of computers. However, these efforts have accelerated as the wealth of scientific literature expands alongside the need to uncover information that is already present in large and unstructured bodies of text, commonly referred to as "non-interactive literatures" (Swanson and Smalheiser, 1997); i.e., literatures that do not cite each other but which, nevertheless, together present useful new information. In addition, the sequencing of the human genome has provided intensive impetus for developing effective tools to identify interrelated concepts and roles such as gene-disease connections and gene-drug interactions from the published literature, as well as from a variety of other types of databases.

This chapter will first explore the importance of biomedical concept relationships and document concept interrelationships, and some of the ways in which concept relationships have been used in information search and retrieval. We will then review a variety of approaches that have been used to represent biomedical concept relationships, beginning with the early concept identification systems developed in the 1950's. Finally, we will look at examples of innovative approaches utilizing biomedical concept identification and relationships for improved document and information retrieval and analysis that support knowledge creation and management.

Before continuing further, a few definitions are in order. A concept can be considered the atom or smallest unit of any knowledge domain or discipline. However, concepts do not exist in isolation; they occur in complex, multidimensional networks that represent "real world" relationships. For the purposes of this chapter, we are using the term "relationship" to denote a semantic association between two or more identified concepts. For example, some typical relationships include: *concept A* "is caused by" *concept B*; *concept A* "is associated with" *concept B*; and *concept A* "is a part of" *concept B*. We will explore the utility of biomedical concept relationships for improving document and information

retrieval and analysis, both within individual documents and among document sets.

Although most of us have a common understanding of the term *relationship*, it is often difficult to explain what appears to be implicit in meaning, even though concept relationships are "…an integral part of the very foundation on which we build and organize our knowledge and understanding of the world in which we live. If concepts are seen as the basic building blocks of conceptual structure, then relationships are the mortar that holds it together" (Green et al., 2001).

The idea that concepts are related to one other is quite useless without knowing the meaning of the relationship. And before a relationship can be identified, "we must be able, first, to designate all the parties bound by the relationship and, second, to specify the nature of the relationship" by identifying the entities that participate in the relationship and the semantics and properties of the relationship (Green, 2001).

Yet, there is a need to be specific and precise when exploring what relationships are, how they are defined and how they can be represented. For example, in the field of mathematics, relational operators such as equals (=), less than (<) or greater than (>) express specific and precise meanings that are well understood by those who are familiar with numbers and mathematics. It would be ideal if our knowledge of relationships in other fields could be interpreted at the same level of precision.

Consider the following scenario: An Alzheimer Disease (AD) researcher is investigating the beneficial effects of caffeine ingestion to slow memory impairment. She knows that caffeine, like adenosine A(2A) receptor antagonists, blocks ß-amyloid–induced neurotoxicity in some rat models for AD. She also knows that caffeine has been shown to improve memory deficits in rat models for Parkinson's Disease (PD). The researcher wants to know if there is a relationship between the protective effects of caffeine consumption and adenosine A2A receptor antagonists for AD patients. She needs to know the level of caffeine dosage ingested over what time period, possible negative and positive associations of caffeine with other neurodegenerative diseases and association of caffeine with other conditions found in an elderly population, such as stroke, high blood pressure, etc. She also wants to know if treatment combining caffeine and adenosine A2A receptor blockers might further slow memory impairment. Using the PubMed®[1] search interface, she searches the MEDLINE® database. Maintained by the National Library of Medicine® (NLM®), MEDLINE

[1] *http://www.ncbi.nlm.nih.gov/entrez/query.fcgi*

contains over 14 million citations to biomedical articles with over 2000 citations added weekly.[2]

The researcher conducts numerous searches using the following terms: *Alzheimer's disease*, *memory*, *caffeine*, *neurodegenerative diseases* and *adenosine A2A receptor antagonists* (hereafter referred to as A2A blockers). Her first PubMed searches on "neurodegenerative diseases" and "Alzheimer's disease" respectively yield 119,829 and 40,785 potential documents. Searching on "A2A blockers" and "caffeine" each retrieve 1209 and 18,965 citations. The researcher then conducts searches combining terms. Combining the terms "memory" and "caffeine" retrieves 173 citations; combining "caffeine" and "A2A blockers" yields 94 citations. Next she tries "neurodegenerative disease" and "caffeine" (93 citations), "neurodegenerative disease" and " A2A blockers " (56 citations) and "Alzheimer's disease" and "caffeine" (21 citations). The researcher then combines "Alzheimer's disease," " A2A blockers " and "caffeine" which yields 9 citations—of which, after careful examination, only 3 appear to present actual answers to her questions.

This scenario includes all of the classic information science problems: precision (a measure of the number of relevant documents as a fraction of all the documents retrieved by the system); recall (a measure of the number of documents useful to the user as a fraction of all the relevant documents retrieved); "aboutness" (subject of the document); and vocabulary control. Traditional search engines and bibliographic database search-and-retrieval systems operate on retrieving a set of documents that reflect only one relationship: a similarity of content matching the keyword or subject terms in the user's query using a basic Boolean keyword retrieval (a query using the Boolean operators "and," "or" and "not"). The implicit interpretation is that there is an equivalent relationship between the user's concept and the document citations retrieved that represent the concept; i.e., the list of citations serves as a surrogate for the requested concept. While this relationship can be presumed and considered useful some of the time, it usually delivers a retrieval set that falls far short of the user's information need and often overwhelms the user with many irrelevant documents.

So the researcher's questions remain unanswered: How much and over what period of time must caffeine be consumed to slow memory loss? Will a combination of caffeine and adenosine A2A receptor antagonists shorten that period of time? Do the neurotoxicity-blocking effects of caffeine and adenosine A2A receptor blockers also come into play with other neurodegenerative diseases? What about possible negative associations with other conditions?

[2] For more information see *http://www.nlm.nih.gov/pubs/factsheets/pubmed.html*

When the researcher entered her terms or term pairs in PubMed, the searching algorithm used the standard keyword approach that counted the words in a query, looked for the presence of terms that matched the query in the database's bibliographic information (i.e., the abstract, title, keywords). While both terms may have been present in the same document, their inclusion in the retrieval set did not necessarily indicate that the two concepts had been studied *in relation* to one another. The term "caffeine" may be present in the abstract as part of an explanation of previous research. The indexer may have included "Alzheimer's disease" in the keywords because it was mentioned in the article introduction. Although the terms "caffeine" and "A2A receptor" are listed as keywords for a document, this does not guarantee that they were studied in relation to one another.

Most approaches to indexing and retrieval of documents to date have not exploited the structure of the document itself as a way of more precisely characterizing biomedical relationships. In addition, most biomedical literature mining has been performed on title and/or abstract words rather than all the words (full-text) in the documents. The researcher's primary information need is to know what was studied and the results of that study. This is data normally located in the Methods and Results sections of a biomedical research report—not necessarily evident in the citation, abstract, subject headings or keywords assigned to the document. Identifying the concept relationships that were studied and reported in the research document is a critical means of matching the biomedical literature user's questions with relevant literature, providing a way to rapidly review, integrate and inter-relate concepts of interest.

2. BACKGROUND AND OVERVIEW: THE USE OF CONCEPT RELATIONSHIPS FOR KNOWLEDGE CREATION

The thesaurus is a key tool developed by information science researchers for displaying the logical, semantic relationships among terms and rules for establishing compilations of terms to denote concepts and concept relationships. Comprised of the specialized vocabulary of a discipline or field of study, the thesaurus is a list of preferred terms to indicate two types of relationships between pairs of terms:

- synonyms, i.e., which of two or more equivalent terms can represent a concept, commonly denoted by Use or UseFor relationships; and
- hierarchical relationships, i.e., broader and narrower terms (parent/child relationships) and association (related terms, such as close siblings).

A more sophisticated type of thesaurus is the ontology, defined as an explicit, formal, systematic specification of all the categories of objects, concepts and other entities in a field or domain; the relations between these categories; and the properties and functions needed to define the objects and specify their actions (see Chapter 8, "Biomedical Ontologies," for a detailed description). Ontologies use rich semantic relationships among terms and strict rules about how to specify terms and relationships.

A key biomedical language resource is the Unified Medical Language System® (UMLS)®, developed by the NLM to overcome information retrieval (IR) problems caused by differences in biomedical terminology.[3] The UMLS consists of three multipurpose knowledge sources that together provide structured representation of concepts and relationships in the biomedical domain:

1. The UMLS Metathesaurus®, a large, multi-purpose, multi-lingual specialized vocabulary database that contains information about biomedical and health-related concepts, their various names and the semantic relationships between concepts.
2. The Semantic Network, a consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts.
3. The SPECIALIST Lexicon, a general English lexicon that includes biomedical vocabulary and a lexical entry that records each term's syntactic, morphological, and orthographic information. The lexical entry is of critical importance to natural language processing (NLP) systems.

One of the more thorough reviews of concept and document relationships occurred at the 1997 ACM/SIGIR workshop "Beyond Word Relations." Participants examined a number of relationship types as possibly significant for IR systems, beyond the traditional topic-matching relationship. The workshop proposed seven relationship types that could prove useful in IR systems:

1. Word-based relationships: documents that share the same vocabulary or word;
2. Attribute-based relationships: relationships based on shared characteristics (e.g., documents *A* and *B* share same author);
3. Document-document hierarchical relationships: situations in which one document is a sub-set or super-type of the other (e.g., document *A* is an appendix or sub-piece of document *B*);
4. Document-document topological relationships: a conceptual extension to the hierarchical relationship, this includes relationships that denote

[3] *http://www.nlm.nih.gov/pubs/factsheets/umls.html*

5. conceptual equivalence (e.g., document *A* is a translation of document *B*), commentary (e.g., document *A* updates document *B*), etc.
6. Document-to-document influence relationships: situations in which one document has affected the writing of another (e.g., document *A* builds on the work of document *B*);
7. Topic-based (or meta-topic based) relationships: this type includes the traditional topic-matching relationship, as well as situations in which documents are related but through less obvious topical resemblances (e.g., "non-interactive literatures" as mentioned earlier);
8. Usage-based relationships: documents that are related through the use of the documents, as in a user profile (Hetzler, 1997).

In addition to issues related to capturing a variety of types of concept relationships, the importance of document structure to indexing strategy, the difficulty of translating a user's information need into a query that can retrieve relevant and useful document representations (whether bibliographic citations or full-text documents) and the problem of how to represent biomedical concept relationships have been under investigation for many years. Our overview will cover various approaches to biomedical literature data mining that focus on utilizing concept relationships, including indexing and vocabulary strategies, information extraction (IE), NLP, text mining and literature-based discovery IR—however, much of the work has narrowly focused in the genome sciences domain. While we provide examples of methods and systems that represent each approach, much more work has been published in this area than can be referenced here.

2.1 Indexing Strategies and Vocabulary Systems

Traditionally, subject access to information has been provided in two ways: (1) classification, the process of describing the subject of an information object (its "aboutness") so it can be uniquely distinguished from all other items and (2) indexing, the process of assigning terms from a controlled vocabulary list (a collection of preferred terms that are used to assist in more precise retrieval of content).

There are two kinds of relationships between terms: semantic and syntactic. Semantic relationships are by definition permanent relationships; they exist independently of document content. For example, the concepts *animal* and *mammal* are related regardless of content. Syntactic relationships, however, consist of otherwise unrelated concepts that are brought together because of the document "space" they share. These relationships are not permanent.

An early innovation in information science research was the coordinate index—a list of terms that can be combined when indexing or searching a

body of literature—which developed into subject-based terminology lists and, more recently, into thesauri and ontologies. In 1952, Taube pioneered the development of IR systems with his invention of the post-coordinate indexing system for subject retrieval. Post-coordinate indexing is the assigning of single concept terms from a controlled vocabulary to a record so that the user is able to "coordinate" or combine the terms using any combination of those concepts in any order when searching (Taube, 1953-57). Post-coordinate index can minimize the number of entries necessary to index all the concepts in a work.

An example of post-coordinate indexing is the MEDLINE database. Documents in the MEDLINE database are indexed using the Medical Subject Headings (MeSH) vocabulary, a post-coordinate indexing strategy. The user can select individual, indexer-assigned concepts from the MeSH controlled vocabulary and combine them with Boolean operators.

A more sophisticated indexing approach to capture biomedical relationships is pre-coordinate indexing, in which terms from the coordinate index are combined at the time of indexing into subject strings that capture concept relationships. Users do not have to coordinate these concepts themselves but can search on the pre-coordinated concepts, resulting in more precise retrieval. For example, MeSH terms can be assigned to a document with one or more of the possible 82 subheadings attached, such as "diagnosis" or "drug therapy." A search on the term "hypertension" with the attached subheading "drug therapy" will retrieve articles on the treatment of hypertension using drugs in a more precise manner than simply connecting the two concepts by an "and" operator, as in post-coordinate systems. The latter could result in retrieval of articles that are about hypertension with the drug therapy directed at another disease.

In coordinate indexing, syntactic relationships are displayed according to the syntax of a normal sentence, either through the syntax of the subject string (precoordinate indexing) or through devices such as facet indicators (postcoordinate indexing). Because of the absence of syntactic relationship indicators in postcoordinate systems, users are unable to distinguish between different contexts for the same term. This can result in retrieval of a set of documents that, while topically related, also contain "false drops" because there is not a mechanism for linking the terms to their respective composite subject or context (Foskett, 1982).

Farradane (1980) proposed Relational Indexing, a framework of nine relationship types, as a scheme for representing structures of syntactic relationships between terms in document descriptions with the goal of providing better retrieval of technical documents. In relational indexing, the meaning in information objects is denoted in the relationships between terms. This approach was not widely utilized, perhaps because the limited

number of relationship types required manual indexing. However, Relational Indexing served as a precursor to some of the features of the UMLS and other thesaurus systems.

Similar to relational indexing, Craven (1978) proposed LIPHIS (Linked Phrase Indexing System), a system of computer-assisted subject indexing that used a network of terms in which arcs correspond to relationships denoted by prepositions. Like Farradane's scheme, the emphasis on concept relationships captured more of the content of an information object than individual term indexing alone.

In many vocabulary systems, conceptual relationships are characterized by generic relationships such as "broader than" and "related to." Other systems, including ontologies, utilize terminologic logic to describe a richer, more informative set of semantic relationships, such as "is_a," "connected_to" and "part_of." In biomedical literature, thesauri concept relationships conform to three general semantic classes of relationships that are used to express various dependencies and connections (Chowdury, 1999):

- Equivalence: which denotes the relationship between a preferred and non-preferred term and is shown through cross-references;
- Hierarchical: which represents pairs of terms in their superordinate (the whole) or subordinate (the part) status and is denoted by "Broader Than" and "Narrower Than" codes; and
- Associative: which describes the relationships between terms that are not in either the hierarchical or equivalence class and is shown by "Related" term codes.

In any natural language text, sequences of characters are combined into words, sequences of words are combined into sentences, sequences of sentences are combined into paragraphs and the sequences of paragraphs into texts. There is, thus, a hierarchy of levels of organization in text and there are corresponding levels of indexing to represent these textual levels. In the case of biomedical research reports, the structure is highly predictable (e.g., introduction, methods and design, research findings and conclusions sections). Approaches to relationship representation described so far have largely ignored the structure of the document in favor of representing isolated concepts. More recent work has focused on systems to extract biomedical relationships in the context of the document structure.

2.2 Integrating Document Structure in Systems

Much of the progress over the last several years in improving text understanding and retrieval has been due to systematic evaluations using complete, naturally-occurring texts as test data conducted at the Message

Understanding Conferences (MUC)[4] and Text REtrieval Conferences (TREC)[5]. MUC and TREC are currently sponsored by the U.S. Advanced Research Projects Agency (ARPA) and have enjoyed the participation of non-U.S. as well as U.S. organizations. MUC focuses on NLP while TREC is IR-focused. Both conferences provide the necessary infrastructure and large test corpora for large-scale, statistically valid performance figures and objective evaluation metrics of NLP and text retrieval methodologies.

Identification and classification of names of person, organization, location, etc., at accuracies exceeding 90% and successful extraction of binary relations among these entities at over 75% accuracy have been reported from these conferences (Aone et al., 1998). Also, as information extraction and retrieval systems have improved, attention has recently been turning to the potential contribution of document structure for text understanding and retrieval.

For example, Yeh et al. (2003) report the results of a Challenge Evaluation task created for the Knowledge Discovery and Data Mining (KDD) Challenge Cup to identify the set of genes discussed in a training corpus of 862 journal articles curated in FlyBase, a comprehensive database for information on the genetics and molecular biology of *Drosophila*. The common feature among the "winning" systems was use of document structure (i.e., concentrating on only certain sections of the document, for example, the "Results" or "Methods" sections and avoiding sections such as "References" in which citations will include names of genes not discussed in the paper) and/or linguistic structure (e.g., sections, paragraphs, sentences, and phrases), as well as table and figure captions, as a means of limiting where to look for features or patterns. The authors note that:

> This is in contrast to the information retrieval approach of treating a paper as just an unstructured set of words. We expect that systems will need to make more extensive use of linguistic and document structure to achieve better results and to accommodate more realistic tasks. (Yeh et al., 2003, pp. i338-9)

A similar approach is used by the PASTA (Protein Active Site Template Acquisition) Project system, which focuses on extracting information concerning the roles of particular amino acid residues in known three-dimensional protein structures. Text preprocessing includes a module that analyzes the text structure to determine which sections will proceed to continued processing. Since certain term classes may occur in only one particular section of text, by leveraging the standard structure of a scientific

[4] *http://www-nlpir.nist.gov/related_projects/muc/*

[5] *http://trec.nist.gov*

article, PASTA can exclude those portions of text that are not of interest. In addition, the PASTA system uses the document section to alter processing (Gaizauskas et al., 2000).

Another system that processes only specific parts of documents is FigSearch (Liu et al., 2004), a classification system that focuses specifically on a document's table and figure legends. The system ranks figures as likely to represent a certain type (e.g., protein interactions, signaling events) and allows users to search for these specialized subsets of figures from full-text.

Although incorporating the document's structure can help reduce the scope of material needing processing and potentially reduce inevitable "noise" in the results, this approach is not without its limitations. A major criticism of these systems is their specialization and consequent difficulty in porting to new domains or use in new applications. Also, the advantages of specialization (e.g., faster processing time) are achieved at the cost of limited terminology handling. For example, terminological issues of synonyms and term variants, expanding abbreviations, and lack of a mechanism for handling relations between terms continue to be problems encountered in systems that incorporate document structure.

2.3 Text Mining Approaches

Text mining refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. Text mining systems generically involve preprocessing document collections using text categorization and term extraction, storing the intermediate representations for analysis (e.g., distribution analysis, clustering, etc.) and visualizing the results. Association rules, which link pairs or larger groups of concepts and are usually assigned support and confidence values, are a dominant analysis method in text mining research. An association rule such as *concept A* → *concept B* indicates there may be a potentially interesting directional association from *A* to *B*. Typically these are discovered by exploiting the co-occurrence of concepts in the texts being mined (Hristovski et al., 2004).

Hirschman et al.,'s (2002) review of milestones in biomedical text mining research, notes that the field began by focusing on three approaches to processing text:

- Linguistic context of the text, such as the work of Fukuda et al. (1998) who pioneered identification of protein names;
- Pattern matching, as seen in the work of Ng and Wong (1999), who used templates that matched specific linguistic structures to recognize and extract protein interaction information from MEDLINE documents; and

- Word co-occurrence, such as Stapley and Benoit (2000) who extracted co-occurrences of gene names from MEDLINE documents and used them to predict their connections based on occurrence statistics.

As data mining technologies and NLP systems improve, more complex text can be processed and corpus-based approaches developed, as seen in the work of Pustejovsky et al. (2002), who used a corpus-based approach to develop rules specific to a class of predicates on a corpus of *inhibit*-relations; and Leroy and Chen (2005), whose Genescene system uses prepositions as entry points into phrases in the text, then fills in a set of templates of patterns of prepositions around verbs and nominalized verbs. NLP has also been used to capture specific relations in databases. For example, EDGAR is a system that extracts relationships between cancer-related drugs and genes from biomedical literature, incorporating a stochastic part of speech tagger, syntactic parser and semantic information from the UMLS (Rindflesch et al., 2000).

These systems have worked to overcome some of the limitations previously mentioned—such as decoding acronyms and abbreviations and detecting synonyms—using machine learning methods, NLP and incorporating ontologies (e.g., the Gene Ontology (GO)[6], a controlled vocabulary of genes and their products).

2.4 Literature-based Discovery IR Systems

As stated in the Introduction, the idea of discovering new relations from a bibliographic database was introduced by Swanson as "undiscovered public knowledge" that merit further investigation. Figure 7-1 illustrates Swanson's characterization of one of his first "mutually isolated literatures": Raynaud's disease, a peripheral circulatory disorder, and dietary fish oil. Although each of the two literatures were public knowledge, they were not bibliographically-related (i.e., did not cite one another), but were linked through intermediate literatures that had not been noticed before (Swanson and Smalheiser, 1997).

The premise of Swanson's approach is, given a body of literature reporting that *concept A* influences or is related to *concept B*; and given another body of literature reporting that *concept B* is related to or influences *concept C*; it may be inferred that *concept A* is linked to *concept C*, and if this relationship has not been experimentally tested, there is the potential to uncover previously "undiscovered public knowledge," form hypotheses, and investigate the relationship between *concept A* and *concept C*.

[6] *http://www.geneontology.org/*

In collaboration with Smalheiser over the past two decades, Swanson explored potential linkages via intermediate topics or specializations between bibliographically disconnected areas of specialization. Using this method, several concept relationships have been discovered and proposed for hypothesis testing, including the relationship between migraine and magnesium (Swanson and Smalheiser, 1997) and automatically identifying viruses that may be used as bioweapons (Swanson et al., 2001), among others. ARROWSMITH, an interactive discovery system based on Swanson and Smalheiser's methods, was created in 1991 and continues development today.[7]

Since the introduction of literature-based discovery, efforts to automate this approach and develop discovery algorithms that can be applied to a knowledge base or bibliographic database have resulted in several systems. One such system is BITOLA (Figure 7-2), which applies a general literature discovery algorithm to a knowledge base derived from the known relations between biomedical concepts (MeSH descriptors plus gene symbols in the document title and abstract fields) in the MEDLINE bibliographic database.

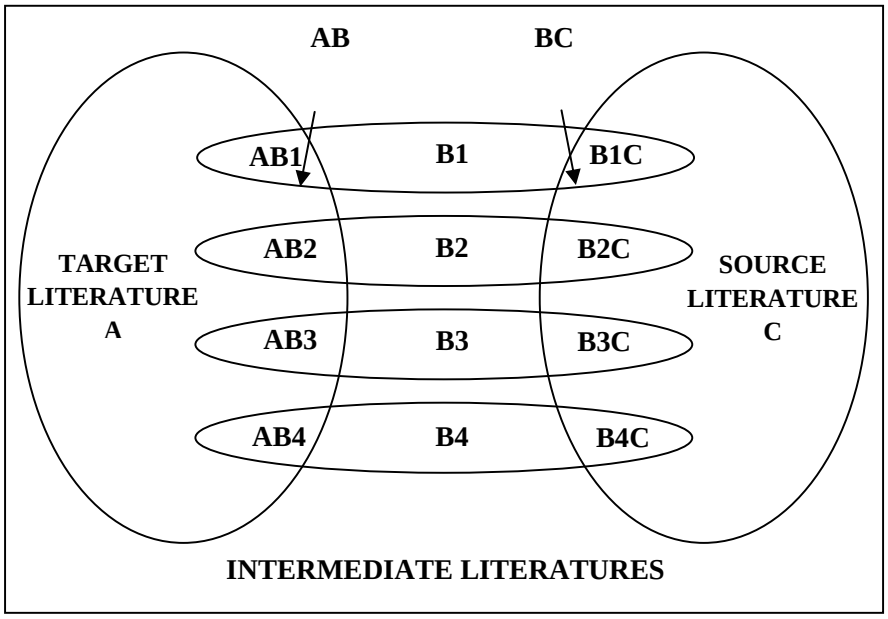


Figure 7-1. A Venn diagram that represents sets of articles or "literatures," *A* and *C*, that have no articles in common but which are linked through intermediate literatures *Bi* (*i* = 1,2,…).

[7] There are two implementations available on the Web: *http://kiwi.uchicago.edu* and *http://arrowsmith.psych.uic.edu*

Such a structure may contain unnoticed useful information that can be inferred by combining pairs of intersections *ABi* and *BiC*. (From: Swanson and Smalheiser, 1997; reproduced with permission of the author.

BITOLA uses HUGO (Human Genome Organisation), the National Center for Biotechnology Information's (NCBI) LocusLink (a database of curated sequence and descriptive information about genetic loci) and OMIM (NCBI's Online Mendelian Inheritance in Man catalog of human genes and genetic disorders) as sources for gene symbols and names as well as gene locations. It also uses OMIM to obtain chromosomal locations. To decrease the number of candidate relations and make the system more suitable for disease candidate gene discovery, the system includes genetic knowledge about the chromosomal location of the starting disease as well as the chromosomal location of the candidate genes (Hristovski et al., 2004).

Similar to Swanson's procedure, BITOLA first finds all the *concepts Y* that are related to the starting *concept X* (e.g., if *X* is a disease then *Y* might be pathological functions). Then all the *concepts Z* related to *concepts Y* are found (e.g., if *Y* is a pathological function, *Z* might be a molecule related to the pathophysiology of *Y*). Finally, the medical literature is searched to check whether *concept X* and *concepts Z* appear together. If they do not appear together, there is the possibility that a new relationship between *concept X* and *concept Z* has been discovered. Figure 7-2 illustrates the BITOLA literature discovery system.

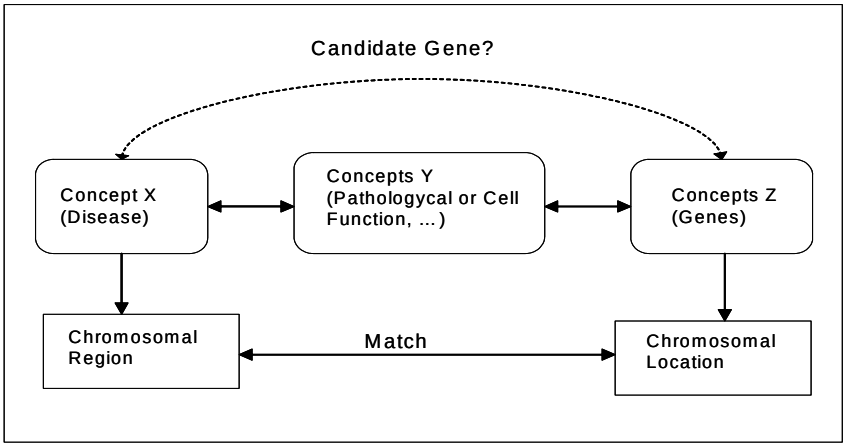


Figure 7-2. Discovery algorithm overview as applied to candidate gene discovery. For a starting disease *X*, we find the related concepts *Y* (disease characteristics) according to the literature (MEDLINE), then find the genes *Z* that are related to disease characteristics *Y*. (From: Hristovski et al., 2004; reproduced with permission of the author.)

As mentioned earlier, one criticism of literature-based discovery IR systems is the limitation imposed by utilizing only titles and/or abstracts. Another criticism, given the exponential growth in MEDLINE records, is their reliance on the use of words to "bridge" between unrelated domains; with new information and records added to MEDLINE on a daily basis, the scale of analysis required by these systems will continue to increase. In addition, other limitations cited include the absence of using synonyms to control vocabulary, the ambiguity produced when abbreviations are not automatically expanded and, particularly with gene symbols, the redundant use of a particular symbol that has differing meanings depending on the context (Wren et al., 2004).

2.5 Summary

Several of the systems already mentioned are, in fact, hybrid systems combining different text mining approaches (NLP, co-occurrence techniques, machine learning, etc.) with incorporation of different knowledge sources (UMLS Metathesaurus, GO, OMIM, etc.). It is obvious that there is not one approach that might be applied to the diverse and wide-ranging biomedical literature. Text formats vary from structured to unstructured and systems vary from free-text analysis to those that focus on document sections (titles and/or abstracts; specific document sections such as methods or table/figure captions). While most of the systems mentioned incorporate some controlled vocabulary component to reduce "noise," efforts to overcome terminologic issues—such as synonym, acronym and abbreviation ambiguity—vary widely.

It is notable, however, that most systems are genomic or proteomic-specific. The issue of scalability of these specialized systems will play an increasing role in their utility and future use as pathways for knowledge creation and discovery.

3. CASE EXAMPLES

Vannevar Bush, often referred to as one of the early pioneers of what later emerged into hypertext systems and the World Wide Web, suggested the use of associations as the main organizing mechanism when filing and retrieving records of information, and described an information space based on the use of associative "trails" to retrieve information (Bush, 1945). We have already mentioned several innovative approaches that are using biomedical concept relationships for improved document and information retrieval and analysis. In this section, we highlight two systems that include

biomedical concept relationship extraction—Genescene and Telemakus—and embody the suggestions of Vannevar Bush in making the increasing body of recorded knowledge more easily accessible.

3.1 Genescene

Genescene,[8] which focuses on cellular processes, utilizes published MEDLINE abstracts and allows retrieval and visualization of biomedical relations extracted from the content of the abstracts. It uses a linguistic relation parser and Concept Space, an automatically-generated, corpus-based co-occurrence thesaurus of semantically related concepts. The system combines bottom-up and top-down approaches. The parser provides precise and semantically rich relations with a rule-based top-down algorithm. Concept Space captures the relations between semantic concepts from large collections of text using bottom-up techniques. The overall system offers a bottom-up view on the data in that the data is allowed to speak for itself, generating interesting patterns or associations that can be used to form new hypotheses. What follows is a summary of the system as described in Leroy and Chen (2005).

The process of creating the Concept Space begins with a download of MEDLINE abstracts in XML format. Abstract and title areas are selected. Using the AZ Noun Phraser (Tolle and Chen, 2000) optimized for biomedical language by using the UMLS SPECIALIST Lexicon as a lexical lookup—the linguistic parser extracts noun phrases. Phrases are analyzed and sorted so that each phrase becomes represented as a concept. Phrase and document frequencies are computed and used to weight each phrase. Concept Space is generated in the final step, co-occurrence analysis. Co-occurrence analysis produces a list of weighted, related noun phrases and their individual components (e.g., modifiers, etc.). Noun phrases are semantically tagged by three ontologies: HUGO, GO and the UMLS. The relations between noun phrases represent the relationships in the entire collection of abstracts originally downloaded from MEDLINE. Figure 7-3 illustrates the Genescene process. Future enhancements include an interactive, graphical map display and visual text mining.

By expanding biomedical literature mining beyond simply identifying genes and proteins and by providing a means for researchers and scientists to discover previously undiscovered gene associations, systems like Genescene will play an increasingly critical role in aiding research, knowledge creation and biomedical discovery.

[8] A demo is currently accessible through *http://ai.eller.arizona.edu/*

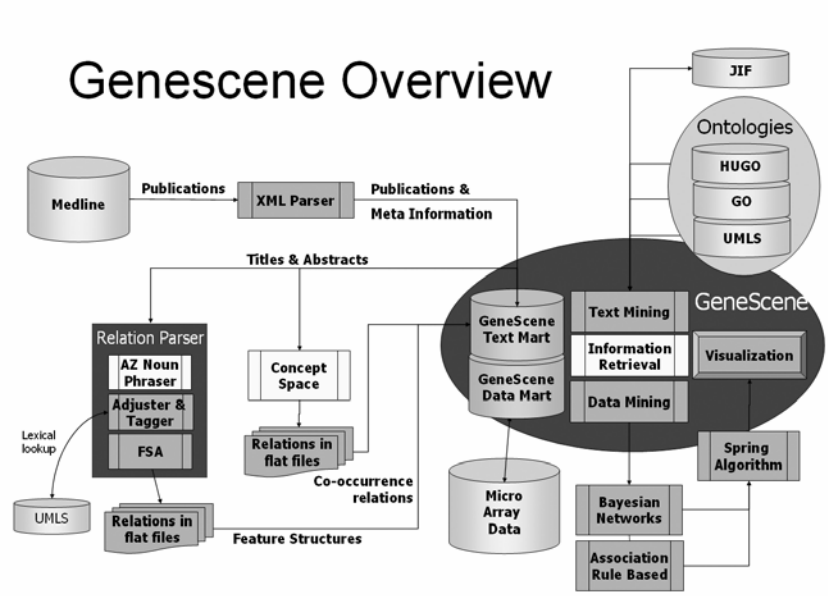


Figure 7-3. Overview of the Genescene architecture Relation Parser. (From: Leroy and Chen, 2005; reproduced with permission of the author.)

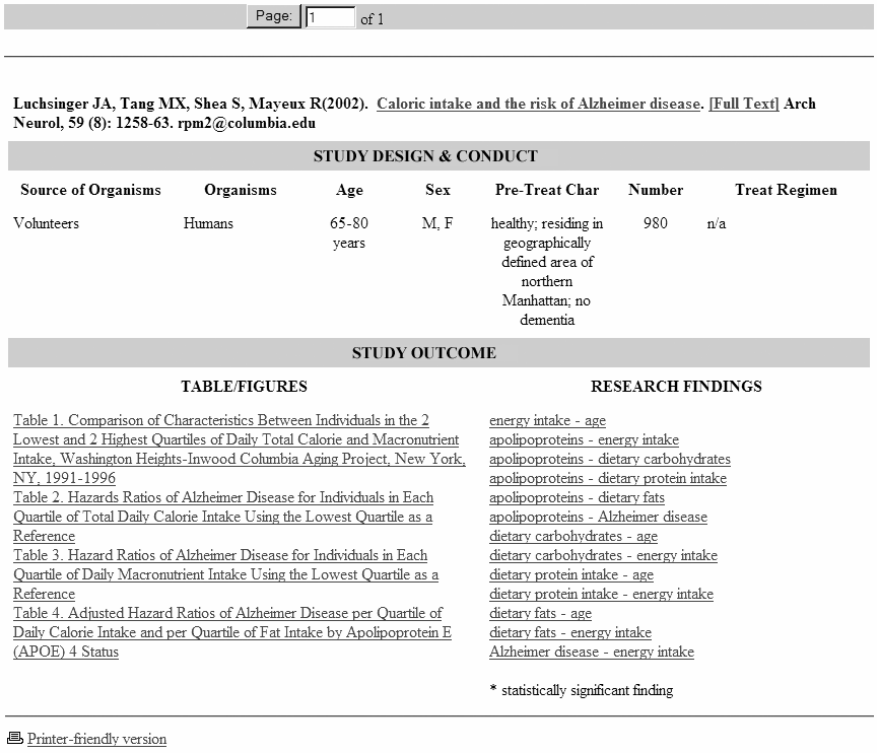
3.2 Telemakus

A hybrid system, Telemakus[9] has a broader focus than any system mentioned other than ARROWSMITH. Although Telemakus currently mines the biomedical literature concerned with the biology of aging, the tools and system architecture have been built to handle other biomedical domains as well. Telemakus uses document structure to limit its extraction of research parameters (e.g., age and number of subjects, treatment, etc.) and concept relationships by focusing on the document's methods section and table and figure captions (Fuller et al., 2004; Revere et al., 2004).

In brief, Telemakus processing is initiated by an analyst who runs, reviews and edits as necessary extractions from the document being processed. The process begins similarly to Genescene, although Telemakus utilizes other bibliographic databases in addition to MEDLINE. The first phase is a download of the document's citation details (in XML format) from which are extracted specific bibliographic fields. The electronic document is then processed to extract additional data, including research parameters and the table and figure legends. These document attributes are loaded into a

[9] Telemakus is freely available at *http:www.telemakus.net/*

database for populating the document conceptual schema—a schematic representation or surrogate of the document with extracted representations of research environment, methods and findings (see Figure 7-4). Telemakus employs the UMLS to control the vocabulary used for some fields, for curating domain-specific thesauri and for concept relationship analysis.



Page: 1 of 1

Luchsinger JA, Tang MX, Shea S, Mayeux R(2002). Caloric intake and the risk of Alzheimer disease. [Full Text] Arch Neurol, 59 (8): 1258-63. rpm2@columbia.edu

STUDY DESIGN & CONDUCT

Source of Organisms	Organisms	Age	Sex	Pre-Treat Char	Number	Treat Regimen
Volunteers	Humans	65-80 years	M, F	healthy; residing in geographically defined area of northern Manhattan; no dementia	980	n/a

STUDY OUTCOME

TABLE/FIGURES	RESEARCH FINDINGS
Table 1. Comparison of Characteristics Between Individuals in the 2 Lowest and 2 Highest Quartiles of Daily Total Calorie and Macronutrient Intake, Washington Heights-Inwood Columbia Aging Project, New York, NY, 1991-1996	energy intake - age apolipoproteins - energy intake apolipoproteins - dietary carbohydrates apolipoproteins - dietary protein intake apolipoproteins - dietary fats apolipoproteins - Alzheimer disease dietary carbohydrates - age dietary carbohydrates - energy intake dietary protein intake - age dietary protein intake - energy intake dietary fats - age dietary fats - energy intake Alzheimer disease - energy intake
Table 2. Hazards Ratios of Alzheimer Disease for Individuals in Each Quartile of Total Daily Calorie Intake Using the Lowest Quartile as a Reference	
Table 3. Hazard Ratios of Alzheimer Disease for Individuals in Each Quartile of Daily Macronutrient Intake Using the Lowest Quartile as a Reference	
Table 4. Adjusted Hazard Ratios of Alzheimer Disease per Quartile of Daily Calorie Intake and per Quartile of Fat Intake by Apolipoprotein E (APOE) 4 Status	

* statistically significant finding

Printer-friendly version

Figure 7-4. Telemakus document schematic representation

The relationship analysis procedure is currently in the process of being automated by incorporating MetaMap (Aronson, 2001), NLM's NLP tool, in combination with term co-occurrence analysis. MetaMap maps arbitrary text to concepts in the UMLS Metathesaurus; or, equivalently, it discovers Metathesaurus concepts in text by parsing text into noun phrases, collecting all UMLS terms containing one or more noun phrases or their variants, and ranking the candidate UMLS terms according to their similarities to all the noun phrases in the text.

Concept identification and assignment of concept relationships to individual documents is derived from processing the document's data tables and figures. Concentrating on data tables and figures focuses the concept

identification and relationship process and reduces the background noise of the full-text document, making the process tractable.

The Telemakus concept relationship approach is of primary significance. On the document level, the basic motivation behind this analysis is to identify what was actually studied and reported. On the larger, domain level, the basic motivation is to capture concept relationships between the documents that form the domain. The interlinking relationships among concepts are represented graphically as concept maps. Concept mapping is a means of spatially representing knowledge in a visual format and, in the Telemakus system, displays the interrelationships between documents and reported research findings.

Visualization of concept relationships offers significant advantages over a textual listing or graph of relationships in that a spatial representation provides a way for users to interact directly with complex information. Visualization and visual exploration can assist in understanding conceptual relationships across a domain and even assist in identifying previously overlooked potential research connections. A strength of concept mapping is that—even though it does not measure strength of relationship between concepts—by aggregating links to concepts as in a many-to-one relationship, a measure of strength is added. In addition, a visualization of concept relationships may be significant for hypothesis generation, as the lack of linkages (the "non-interactive literatures") is more visually apparent.

The following section demonstrates how an information system that incorporates concept relationships can support knowledge creation and discovery.

3.3 How Can a Concept Relationship System Help with the Researcher's Problem and Questions?

Returning to the scenario at the beginning of this chapter, the AD researcher has a number of questions regarding the potential for treatment of the cognitive disabilities of AD or other neurodegenerative diseases with caffeine and adenosine A2A receptor antagonists (A2A blockers):

- Both caffeine and A2A blockers are reported to have neurotoxicity-blocking effects. Is there any relationship between caffeine consumption and adenosine A2A receptor antagonists?
- Adenosine A2A receptor antagonists have been successfully used to treat Parkinson's Disease. Have they been used to treat AD?
- Has anyone studied the relationship between caffeine and memory loss in AD patients? If yes, how much and over what period of time must caffeine be consumed to slow memory loss? Will a combination of

caffeine and adenosine A2A receptor antagonists shorten that period of time?

- Do the neurotoxicity-blocking effects of caffeine and adenosine A2A receptor antagonists also come into play with other neurodegenerative diseases? What about possible negative associations with other conditions?

Another way to look at these questions is in terms of the concept relationships they represent as listed in Table 7-1. It is notable that not one concept relationship can encapsulate the researcher's information need and that some individual concepts can be related to multiple concepts.

Table 7-1. List of Concepts and Possible Concept Relationships

Concept 1	Concept 2	Possible Semantic Relationship(s)
caffeine	memory loss	associated_with / co-occurs_with
memory loss	Alzheimer's disease	manifestation_of
caffeine	neurodegenerative diseases	treats
caffeine	high blood pressure	associated_with / complicates
caffeine	neurotoxicity-blocking effects	associated_with
caffeine	A2A blockers	associated_with
A2A blockers	neurotoxicity-blocking effects	associated_with
A2A blockers	Parkinson's Disease	treats

As mentioned previously, some of the information the researcher needs will be found in the methods or results sections or in the figures and tables rather than the document's title or abstract. For an information need such as this, a system like Telemakus is an appropriate resource with its extracted representations of the research environment, methods and outcomes of the retrieved documents.

Providing the schematic representations (schemas) of retrieved documents allows the researcher to "browse" the document retrieval space without needing to read the articles in their entirety. In addition, characterizing the concept relationships from each document in a visual format maintains the inter-relationships between documents and reported research findings, as well as assists in understanding conceptual relationships across a domain.

Returning to our scenario, when the AD researcher uses Telemakus for her query, she enters terms similarly to the approach she used with PubMed. However, the list of citations returned provides a pathway to both the content of each document—its methods and research findings—and to interactive concept maps of linked relationships across the group of research reports. Figure 7-5 illustrates the retrieval set interface for a search on

Alzheimer's disease and caffeine. From this list, the researcher can browse the schematic representations or schemas of the content of each document (Figure 7-6).

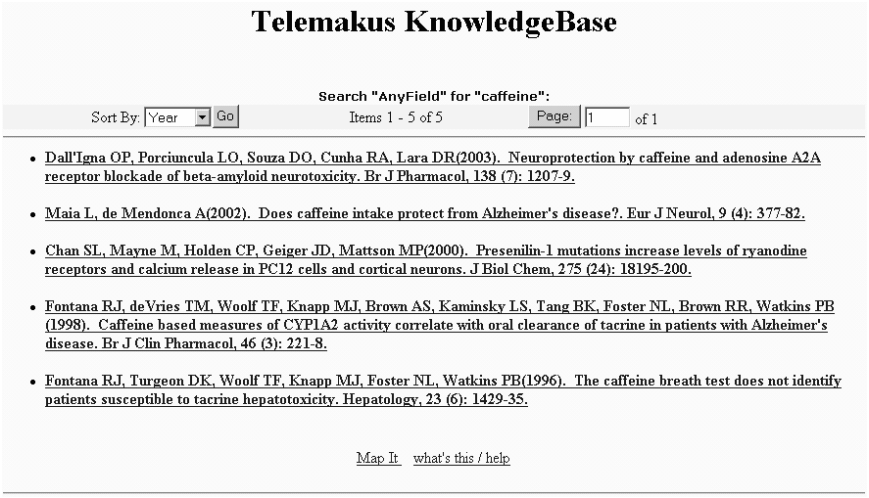


Figure 7-5. Caffeine and AD search retrieval set

Fontana RJ, deVries TM, Woolf TF, Knapp MJ, Brown AS, Kaminsky LS, Tang BK, Foster NL, Brown RR, Watkins PB (1998). Caffeine based measures of CYP1A2 activity correlate with oral clearance of tacrine in patients with Alzheimer's disease. [Full Text] **Br J Clin Pharmacol, 46 (3): 221-8. Dept of Internal Medicine, Univ of Michigan, Ann Arbor, MI 48109**

STUDY DESIGN & CONDUCT

Source of Organisms	Organisms	Age	Sex	Pre-Treat Char	Number	Treat Regimen
recruitment	aged; Aged, 80 and over	>50 years	M, F	mild to moderate Alzheimer's disease, non-smoking, no significant active medical problems	19	caffeine or tacrine administration

STUDY OUTCOME

TABLE/FIGURES	RESEARCH FINDINGS
Fig 1. Individual plasma tacrine concentrations after oral administration of 40 mg tacrine. Table 1. Correlation between Cytochrome P-1A2 activity measures and tacrine pharmacokinetic parameters. Table 2. Pharmacokinetic parameters of parent tacrine and 1-OH tacrine in 19 patients Fig 2. Individual oral clearance of tacrine and estimates of Cytochrome P-1A2. a: Caffeine Breath Test; b: Caffeine metabolic ratio; c: Paraxanthine/caffeine urinary metabolite ratio. Fig 3. Individual subject correlationsof tacrine and 1-Oh tacrine area under the curve values.	tacrine - caffeine cytochrome P-450 - tacrine phosphodiesterase inhibitors - tacrine Alzheimer disease - tacrine Alzheimer disease - caffeine Alzheimer disease - cytochrome P-450 * statistically significant finding

Figure 7-6. Schematic representation of a document from searching "caffeine" and "AD" schema

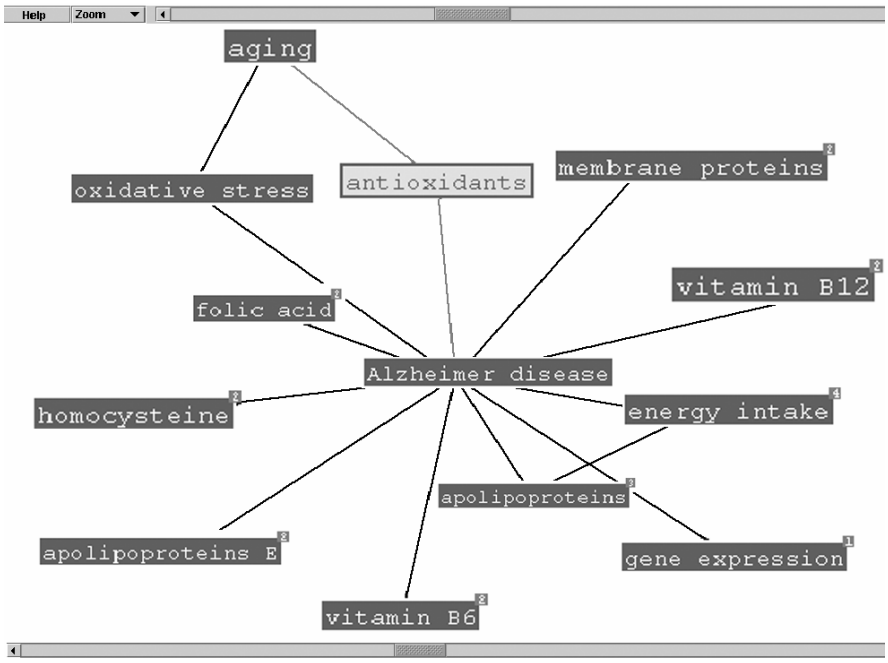


Figure 7-7. Telemakus concept map generated from a search on Alzheimer's disease. Note that edges (lines) of the map signify relationships between concepts, but length does not reflect any weighting scheme.

From the list of retrieved citations, the researcher can also activate the concept mapping function for access to a visualized map of concept relationships for the current retrieval set (Figure 7-7) by selecting "Map It" (at the bottom of the retrieval list in Figure 7-6).

Within each schema, she can access the abstract, document full-text and, by clicking on any blue highlighted item under "Table/Figure," she can link to each table or figure in the full-text article. It is from the schema that the researcher will obtain answers to several of her questions by either examining a table or figure of interest or by browsing the information found under "Study Design and Conduct."

The researcher can also explore individual concept relationships by selecting any pair listed under "Research Findings" in the schema, activating a search of the relationship across all documents in the Telemakus knowledgebase. For example, selecting the pair "caffeine – cell death" will result in a retrieval set listing all documents to which this concept relationship has been assigned. The concept maps generated by Telemakus will help the researcher answer other questions.

Further exploration can be done within the concept map by selecting linkages between concept relationships and by selecting individual concepts.

The iterative nature of the search process and ability to explore research connections from both the schematic and concept map interfaces can support the process of knowledge discovery in a way that mimics the way many scientists work—by providing a means of exploring a variety of types of connections and potentially discovering a new frame of reference for the information problem.

3.4 Summary

Literature-based discovery systems provide the potential for supporting "systematic serendipity." Originally coined by Garfield, systematic serendipity refers to the organized process of discovering previously unknown scientific relations using citation databases, leading to better possibilities for a collaboration of human serendipity with computer-supported knowledge discovery.

Hypothesis generation and testing are critical steps towards making scientific discoveries. Along with the "aha" experience of insight and discovery, hypothesis generation requires prior knowledge. Yet researchers are sometimes unaware of relevant work by others that could be integrated into theirs or are unable to put together enough "pieces" of their domain's jigsaw puzzle to recognize that one is missing or being overlooked. As illustrated in the scenario above, literature-based knowledge discovery systems that include concept relationships and schematic representations are a means of providing additional pathways to these puzzle pieces.

4. CONCLUSIONS AND DISCUSSION

This chapter has presented a variety of approaches that have been used to characterize biomedical concept relationships and document concept interrelationships and some of the ways in which concept relationships have been used in information search and retrieval. We have reviewed a very small number of the innovative approaches utilizing biomedical concept identification and relationships for improved document and information retrieval and analysis. This arena in the knowledge discovery field—utilizing biomedical concept relationships—is fairly young and promises to be a rich and interdisciplinary endeavor.

In our survey of approaches to indexing and retrieval of documents, we have noted that most systems rely on title and/or abstract text to assign or mine concept relationships. In addition, most systems do not exploit the structure of the document itself as a way of more precisely characterizing

biomedical relationships. As more and more biomedical literature becomes available electronically, we will likely see an increase in extraction approaches that incorporate the full-text document.

We have assumed, given the rapid expansion of scientific research, that there is great need for creating systems that aid in finding or integrating new domain knowledge. While we have focused on the concept relationship component of knowledge management systems that support biomedical research, we have not thoroughly discussed the role and realistic utility of such systems for creating knowledge. A significant research area that requires attention is evaluation of these systems. Usability research is needed to validate the utility of these approaches for scientists and researchers. There are numerous questions for literature-based concept relationship systems, including the following:

How exactly does including concept relationships in such systems support the creative process of hypothesis development?

How can systems avoid hindering the "eureka" experience of scientific research?

What methods must be employed to answer these questions?

While Swanson and Smalheiser established a discovery framework that has been used by many researchers as a measurement standard, extensive and comprehensive evaluation efforts are still needed for validating the contribution these systems can make for knowledge creation and discovery.

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge the helpful suggestions made by anonymous reviewers.

REFERENCES

Aone, C., Halverson, L., Hampton, T. and Ramos-Santacruz, M. SRA (1998). "Description of the IE2 System Used for MUC-7," in *Proc 7th Message Understanding Conference*.

Aronson, A. R. (2001). "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program," in *Proceedings of the AMIA Annual Fall Symposium*, pp. 17-21.

Bush, V. (1945). "As We May Think," *The Atlantic Monthly*, 176*(1)*:101-108.

Chowdury C. G. (1999). Introduction to Modern Information Retrieval. Library Association Publishing.

Craven, T. C. (1978). Linked Phrase Indexing. *Information Processing and Management*, 14(6):469-76.

Farradane, J. (1980). "Relational Indexing, Parts I and II," *Journal of Information Science*, 1(5-6):267-76; 313-24.

Foskett A. C. (1982). *The Subject Approach to Information*. 4th ed. Hamden, CT: Linnet Books.

Fukuda, K., Tamura, A., Tsunoda, T. and Takagi, T. (1988). "Toward Information Extraction: Identifying Protein Names from Biological Papers," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 707-18.

Fuller, S., Revere, D., Bugni, P. and Martin, G.M. (2004). "A Knowledgebase Information System to Enhance Scientific Discovery: Telemakus," *BMC Digital Libraries*;1*(1)*:2 (21 September 2004).

Gaizauskas, R., Demetriou, G. and Humphreys, K. (2000). "Term Recognition and Classification in Biological Science Journal Articles," in *Proceedings of the Computational Terminology Workshop for Medical and Biological Applications*, pp. 37-44.

Garfield, E. (1966). "The Who and Why of ISI," *Essays of an Information Scientist*, 13:33-37.

Green, R., Bean, C. A. and Myaeng, S. H. (2002). "Introduction," in R. Green, C.A. Bean and S.H. Myaeng (eds.), *The Semantics of Relationships: An Interdisciplinary Perspective*, pp. vii-xviii. Kluwer Academic Publishers.

Green R. (2001). "Relationships in the Organization of Knowledge: An Overview," In C.A. Bean and R. Green (eds.), *Relationships in the Organization of Knowledge*, pp. 3-18. Kluwer Academic Publishers.

Hetzler, B. (1997). "Beyond Word Relations: SIGIR '97 Workshop," *ACM SIGIR Forum,* 31(2):28-33.

Hirschman, L., Park, J. C., Tsujii, J., Wong, L. and Wu, C. H. (2002). "Accomplishments and Challenges in Literature Data Mining for Biology," *Bioinformatics*, 18(12):1553-61.

Hristovski, D., Peterlin, B., Mitchell, J. A. and Humphrey, S. M. (2004; in press). "Using Literature-based Discovery to Identify Disease Candidate Genes," *International Journal of Medical Informatics*.

Leroy, G. and Chen, H. (2005; in press). "Genescene: An Ontology-enhanced Integration of Linguistic and Co-occurrence Based Relations in Biomedical Texts," *Journal of the American Society for Information Science and Technology*.

Liu, F., Jenssen, T.-K, Nygaard, V., Sack, J. and Hovig, E. (2004; in press). "FigSearch: A Figure Legend Indexing and Classification System," *Bioinformatics*.

Ng, S. K. and Wong, M. (1999). "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts," in *Proceedings of the Genome Informatics Series: Workshop on Genome Informatics*, pp. 104-112.

Pustejovsky, J., Castano, J., Zhang, J., Kotecki, M and Cochran, B. (2002). "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 362-73.

Revere, D., Fuller, S.S., Bugni, P. and Martin, G.M. (2004). "An Information Extraction and Representation System for Rapid Review of the Biomedical Literature," Accepted for Presentation: *MedInfo*, Sept 2004, San Francisco, CA.

Rindflesch, T. C., Tanabe, L., Weinstein, J. N. and Hunter, L. (2002). "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," in *Proceedings of the Pacific Symposium on Biocomputing, p*p. 517-28.

Stapley, B. J. and Benoit, G. (2002). "Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in Medline Abstracts," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 529-40.

Swanson, D. R. and Smalheiser, N. R. (1997). "An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery," *Artificial Intelligence*, 91*(2)*:183-203.