Journal of the American Society for
Information Science and Technology

# Recent Advances in Literature Based Discovery

| | |
|---|---|
| Journal: | *Journal of the American Society for Information Science and Technology* |
| Manuscript ID: | JASIST-2005-09-0156 |
| Wiley - Manuscript type: | Research Article |
| Date Submitted by the Author: | 20-Sep-2005 |
| Keywords: | knowledge discovery < automatic extracting < computer applications < computer operations < (activities and operations), data mining < knowledge discovery < automatic extracting < computer applications < computer operations < (activities and operations), machine learning < artificial intelligence < computer applications < computer operations < (activities and operations) |
| | |

# Recent Advances in Literature Based Discovery

**Murat C. Ganiz, William M. Pottenger and Christopher D. Janneck**

{mug3, billp, cdj2}@lehigh.edu

Computer Science & Engineering Dept., Lehigh University,

19 Memorial Drive W., Bethlehem, PA 18015

Correspondence to:

Murat C. Ganiz

Computer Science & Engineering Dept., Lehigh University,

19 Memorial Drive W., Bethlehem, PA 18015

Tel: (610) 758-3737; Fax: (610) 758-4096

Email: mug3@lehigh.edu

## Abstract

Literature Based Discovery (LBD) is a process that searches for hidden and important connections among information embedded in published literature. Employing techniques from Information Retrieval and Natural Language Processing, LBD has potential for widespread application yet is currently implemented primarily in the medical domain. This article examines several published LBD systems, comparing their descriptions of domain and input data, techniques to locate important concepts from text, models of discovery, experimental results, visualizations, and evaluation of the results. Since there is no comprehensive "gold standard," or consistent formal evaluation methodology for LBD systems, the development and usage of effective metrics for such systems is also discussed, providing several options. Also, since LBD is currently often time-intensive, requiring human input at one or more points, a fully-automated system will enhance the efficiency of the process. Therefore, this article considers methods for automated systems based on data mining.

## 1. Introduction and Background

In a world with seemingly boundless increases in scientific knowledge, researchers struggle to maintain expertise and knowledge of developments in their fields. More scientific journals, with a greater number of articles per journal, expand already enormous bibliographic databases. One example, the online Medline database, (focusing on biomedicine) contains approximately 13 million references to articles from about 4,800 journals worldwide. Additionally, since 2002, between 1,500-3,500 completed references are added each day, Tuesday through Saturday[1]. As a result, individual scientists must interpret massive amounts of existing knowledge while keeping up with the latest developments.

Dealing with the substantial amount of information has led to a fragmentation of scientific literature. Fragmentation exists within: (1) specialties: e.g., advances in biophysics, astrophysics, mathematical physics; (2) sub-specialties: e.g., aquatic toxicology, proteomics, molecular immunology; (3) structure: e.g., blood, cell, lipids, nucleic acids research; and (4) technique: e.g., electrophoresis, mass spectrometry, ultra microscopy. Swanson (2001) asserts that this specialization or fragmentation of literature is a growing problem in science, particularly in biomedicine. Scientists tend to correspond more within their fragments than with the field's broader community, promoting poor communication between specialties (Swanson & Smalheiser, 1997). This is evidenced within the citations of such literature, as authors exhibit tendencies to heavily cite those within their narrow specialties. Results are published, yet

[1] National Library of Medicine (NLM), MEDLINE Fact Sheet, 2005.
http://www.nlm.nih.gov/pubs/factsheets/medline.html

researchers may never be aware of others' relevant work. This results in interesting and useful connections between fragmented information, though implicit, going unnoticed.

Classical techniques, like conventional computer-aided literature searching or Information Retrieval, are insufficient for recognizing connections. A solution to the problem is Literature Based Discovery (LBD) which directly addresses the problems of knowledge overspecialization. LBD strives to find connections that are novel: that have not been previously explicitly published.

Don R. Swanson (1986) first introduced the concept of discovering new relations within a bibliographic database. He has subsequently contributed several discoveries that have been published in various journals (Swanson, 1986; 1988; 1990, Smalheiser & Swanson, 1994; 1996a; 1996b; 1998a). He defines LBD as the process used to find complementary structures in disjoint science literature. His complementary structures derive from two separate arguments which, when combined, yield novel and important inferences and insights. These arguments are defined as "disjoint" as they are taken from articles that do not mention, cite or co-cite each other. For the purposes of this review, LBD will be characterized by four main points:

1. LBD uses existing knowledge from published science literature (e.g., Medline, or the World Wide Web);
2. The LBD process strives to find connections between two arguments (e.g., "high blood viscosity" and "platelet aggregation" are mentioned in both Fish Oil and Raynaud's Disease literature);

3. The combination of two arguments may yield a new insight that was not originally apparent (e.g., Raynaud's Disease can be treated with Fish Oil); and
4. Any connections made should be novel and previously unpublished (e.g., no article in the Medline database ever mentioned both Fish Oil and Raynaud's Disease together).

**Figure 1: Swanson's discovery: Raynaud's↔Fish Oil connection from (Weeber, 2001)[2]**

There are currently several fundamental complexities with LBD. First, the overall scope is infinite *prima facie*, in that there is a seemingly unmanageable information space with many potential connections. Second, information is represented in an unstructured format – natural language. Third, there is no standardized vocabulary by which to formally define different LBD techniques. Therefore, for the purposes of this work, a standardized phraseology for comparing and understanding different LBD techniques is established, with many of these terms taken from and based on Swanson's framework. These standard terms are summarized in Figure 2 and are further defined below.

**Figure 2: Swanson's framework, which is followed by many other works (Open Discovery)**

The LBD process is comprised of two types of entities, concept and literature One definition of concept is a mental picture of a group of things that have common characteristics. In this work,

[2] Note that in different works disease can be named as A and substance can be named as C. Search can either start from substance or disease.

**concept** is defined as a word or phrase which describes a meaningful subject within a particular field. For example, in the medical field, “migraine” is a concept as it describes a meaningful subject in this area. **Literature** is heretofore defined as a set of documents about a subject, that are generally scientific documents including journal articles and conference papers. Note that there may be different levels of granularity present: one can refer to the general object, “medical literature,” or can specify with the term “migraine literature.” In all cases a meaningful set of documents is described. When referring to “migraine literature,” it is understood that this refers to a set of related documents that address or mention “migraines.” As such, literature often includes a set of documents that include a particular concept.

Most LBD works utilize either open or closed models of discovery. **Open discovery** is characterized by the generation of a hypothesis as a result of the LBD process (Weeber et al., 2001). The exploration starts with a topic of a scientific problem or research question, deemed the **start concept** or **A-concept**. One such example is the topic of disease in the medical domain. Next, this concept is used as a query to an online database (such as Medline) and all documents that include this concept are downloaded. This document set is the **start literature** and labeled as **“A”.** Important terms or phrases can be extracted and processed using several different techniques along with the help of human expert interaction. Each word or phrase in the resulting filtered list is called an **intermediate concept** or **B-concept**, and represents important subjects in the start literature. For example, if *A* is a disease (such as Raynaud’s Disease), then related *B-concepts* can be characteristics or symptoms of the disease (such as “blood viscosity” and “platelet aggregation”). After again consulting the online database, now using the *B-concepts*, the resulting document set is the **intermediate literature** labeled as **“B”**. The intermediate literature is then processed and terms or phrases are extracted again. After a successful LBD process, one concept is selected by a human expert as the **target concept** or **C-concept**. Finally, one can form a hypothesis that concept *C* can be used for treatment of disease *A* via *B*. Continuing the example, Fish Oil can be used to treat Raynaud's Disease because it lowers blood viscosity. The hypothesis is then checked to make sure there is no overlap between the start and target literature.

This same terminology will be used in describing **closed discovery**, which starts with known *A-* and *C-concepts*. The connection between *A* and *C* may be an observed association or a previously generated hypothesis. The new knowledge gained while employing closed discovery involves finding novel *B*'s that explain the initial observation or hypothesis (Weeber et al., 2001).

**Figure 3: Open Discovery vs. Closed Discovery. Adapted from (Weeber et al., 2001). "The solid arrows indicate potentially interesting pathways of discovery, the dashed ones unsuccessful pathways."**

The remainder of this article is organized as follows: Section 2 explores twelve published LBD works and their additional ventures. The reviews include descriptions of input data, domain, statistical and linguistic approaches to extract tokens from text, the use of open and/or closed models of discovery, experimental results, evaluation and visualization of the results (for more recent works). This examination starts with Swanson's approach to LBD, and his initial ideas will be continually reapplied throughout the survey. Section 3 discusses trends in examined LBD works. Section 4 draws attention to the need for underlying theory to develop effective metrics. Section 5 addresses the lack of formal evaluation methods and “gold” standards. Section 6 explores some questions and directions on the future of LBD including: additional approaches for integrating and mining biomedical literature, the tradeoff of greater automation and diminishing human input, and the feasibility of fully automated LBD systems. Possible future applications of LBD are discussed in Section 7 and conclusions are presented in Section 8. A summarized list of all works surveyed is given in Table 1, at the end of the article.

## 2. Overview of Literature Based Discovery

Swanson first introduced the idea of discovering new relations from a bibliographic database, and has made several medical discoveries since 1986, leading to publications in relevant medical journals (Swanson, 1986; 1988; 1990). Swanson’s work demonstrates the existence of pairs of knowledge which, when combined, can directly contribute to solutions of specific medical problems.

In his first discoveries, Swanson performed extensive manual searches in literature databases, reading many titles and abstracts of scientific publications. Since then, several works have contributed more advanced and automated methods for LBD. These methods are the focus of this survey, and will be discussed in this section.

Most work in this area uses Medline as the literature database and employs different techniques. They focus on replicating Swanson’s results or using his results to evaluate their own. One such method is Vos’s model of discovery which focuses on drug profiles interacting with disease profiles (Rikken & Vos, 1995; Weeber et al., 2000). In this model, intermediate concepts (*B-concepts*) may be adverse drug reactions, such as in the DAD-system (which is an acronym for both ‘Drug–Adverse drug reactions–Disease’ and ‘Disease–Adverse drug reactions–Drug,’ depending on its use). Weeber (2004) states that Vos’s model can be considered as a specification of Swanson’s general model in a drug discovery context. The characteristics of the profiles in Vos’s model are the intermediate *B-concepts* in Swanson’s model. The profile for drug “A,” for instance, may include the therapeutic characteristic (B-concept) of “reduction of oxygen demand” whereas “increase of oxygen demand” may be a characteristic of disease “C” (per Figure 4).

**Figure 4: “Vos’s and Swanson’s model of discovery combined.” (Weeber, 2004)**

There is some research that applies LBD outside of the medical literature domain: Valdes-Perez (1994) uses chemistry databases, and Cory (1997) uses humanities databases. Cory (1997) found indirect links between a 20th century poet and an ancient philosopher. Gordon, Lindsay & Fan (2001) and others attempt to see how LBD techniques might be applied to problems in scientific research as well, but using non-traditional data sources. Using the World Wide Web to find connections between different bodies of information, their approach originates with a technology and then uses prominent characteristics in lexical statistics and analysis to find new application areas of the technology. The discovery of novel (or previously unpublished) knowledge is not the goal of all LBD researchers, however. Lindsay and Gordon (1999) discuss the importance of LBD approaches in promoting awareness of existing connections between disjoint literature, and not necessarily the discovery of new links.

### 2.1. Swanson's first discoveries and early work

The purpose of Swanson's pioneering work (Swanson, 1986; 1988; 1990) is to find links that are not previously known by researchers within published information.  A summary of the procedures is given below, and these methods are explained in more detail in Swanson (1991) and Swanson & Smalheiser (1997).

Swanson (1986) first finds a connection between disjoint literature areas linking Fish Oil and Raynaud's syndrome.  According to (Weeber et al., 2001), this first discovery is a coincidence, as he had been reading these distinct articles for different purposes.  Swanson had been studying the literature on Raynaud's Disease (the *C*)[3].  From this, he learned that patients with this disease have a relatively high blood viscosity and increased platelet aggregation function, and are characterized by certain vasoreactive phenomena.  These characteristics are *B-concepts*, and together they form **BC-knowledge**.  Swanson also found that Fish Oil (an *A-concept*), and its active ingredient eicosapentaenoic acid (EPA), lowered blood viscosity and platelet aggregation (Weeber et al., 2001).  Combining this knowledge, he hypothesized that the active ingredients of Fish Oil (omega-3 fatty acids, including EPA) may help Raynaud's patients.  With this hypothesis, he studied the literatures both on Fish Oil and on Raynaud's disease and found no overlap. Therefore, he published his findings in 1986, bringing this connection to light in the medical domain.

[3] Note however that in most of the research reviewed for this survey, Raynaud's (the start literature) is labeled as A, and fish oil (the target literature) is labeled as C.

Gordon & Lindsay (1996) gives the following step-by-step description of Swanson's (open) process:

1. Pick a topic of interest (Raynaud's Disease)
2. Search to find literature C={Raynaud's}[4]
3. Guess that B (e.g., blood factors) should be studied in relation to Raynaud's
4. Search literature C'=C ∩ {blood}
5. Notice two common descriptors: blood viscosity, red blood cell rigidity
6. Search literature A={blood viscosity}⋃{ red blood cell rigidity }
7. Notice the term "Fish Oil"
8. Search literature A= {Fish Oil}
9. Show {Fish Oil}∩{Raynaud's}= { }
10. Show plausible connection between Raynaud's and Fish Oil

Using this same model, Swanson published a second hypothesis, a connection between migraines and magnesium, by examining the titles from the search results in the disjoint literature. He found several intermediate medical terms, including "epilepsy" and "calcium channel blockers," which occur frequently in the titles of both literature sets. He concluded that magnesium insufficiency is involved in migraines, generating this hypothesis by combining two disjoint literatures via that particular connection. According to Swanson (1988) there were no papers discussing this link before eleven indirect connections that he found in the literature. In his later

[4] In this paper, "{X}" signifies "the set of documents on the topic X." {Raynaud's} therefore refers to the set of all documents that mentions "Raynaud's" in at least one of the Medline fields examined on that topic.

work, Swanson (1991) developed a method to generate a hypothesis and test it using the Migraine↔Magnesium connection.

**Domain and Input Data**

Swanson's early works are all in the medical domain and his first discoveries are published in Biomedicine journals. These discoveries, such as Raynaud's↔Fish Oil (Swanson, 1986) or Migraine↔Magnesium (Swanson, 1988), are links between diseases and substances which can be used to treat them.

In his later work (Swanson, 1991), he uses the titles of a subset of migraine literature (consisting of 65 articles) and of a magnesium subset (63 articles) to test the hypothesis based on the 11 connections he discovered between Migraine↔Magnesium.

**Approach and Model of Discovery**

The model of discovery in Swanson's early work (Swanson, 1986; 1988; 1990) is the open model of discovery, and performed mostly manually. These discoveries are primarily made by personally reading related literature and by applying a two-step approach to the discovery process. First, Swanson forms a hypothesis by either coincidence or some other method. The second step validates the hypothesis by extensive literature analysis. In his later work (Swanson, 1991), he uses a closed model as he assumes the target literature has already been identified and his purpose is to find relevant logical connections between start literature (on "migraine") and target literature (on "magnesium"). He uses co-occurrence analysis (for title words and phrases) to provide a partially automated process to guide the human expert in the search process, in this case using the Medline database. First, a list is built from the words that co-occur with "migraine" within titles. Then, all Medline records that contain words from this list (but not "migraine") are identified. Finally, the title words that co-occur with the words in the first list are identified. This process can be summarized in steps derived from (Swanson, 1991):

1. The first list (of *A-concepts*) was formed in a manual process by selecting physiologically significant key words[5], short phrases, and meaningful word combinations from 1,000 downloaded titles containing the word "migraine".
2. Stop words and "obviously unsuitable words" (according to the human expert) were eliminated.
3. Selected words and phrases were organized into about 160 groups by "bringing together and logically combining related terms including synonyms and inflectional variants." Each group was defined by a term or concept.
4. The ratio of frequency of occurrence within migraine titles to frequency in all titles in the Medline database was computed to rank each of the 160 terms. A cut-off ratio of 8 was selected because this led to the identification of 17 (out of the 84 highest-ranking) terms associated with 7 of the 11 known Migraine↔Magnesium connections.

A second Medline search was conducted using these 84 terms, resulting in about 120,000 records. There were only 127 occurrences of the term "magnesium" which is not more significant than random. Therefore, Swanson needed to restrict the search space: with help of the Medline subheadings, he narrowed the target set to deficiency states, dietary factors, poisons and toxicity studies, to find the term "magnesium."

[5] This relies completely on human expert knowledge.

This early work led Swanson to develop a more automated discovery support tool named Arrowsmith (Section 2.2 below). With this tool, users can search for novel connections between two literature sets using the closed approach, which aids medical experts in using Literature Based Discovery.

**Evaluation and Discussion**

Swanson's discoveries such as Raynaud's↔Fish Oil (Swanson, 1986) or Migraine↔Magnesium (Swanson, 1988) are verified by subsequent papers in the biomedicine area. According to Swanson (1991), the main problem in building a fully automatic system is that neither the syntactic nor the semantic structures offer a straightforward chain of deductive reasoning with respect to a possible connection. There is a need for "background knowledge." That knowledge comes from human experts that understand each title's meaning. In the conclusion of this work, Swanson noted that there is some exploitable information in word frequency distributions, but not enough to encourage the attempt to develop a fully automatic process. This background knowledge problem is often mentioned in the following works as the main problem facing the development of fully automated systems, because human judgment must therefore be applied at several points. Neither of these works discuss the potential for a supervised machine learning algorithm to be applied to this problem, which can simulate the human expert by learning background knowledge required in this kind of process. A detailed discussion of this possibility is contained in Section 5.

### 2.2. Swanson & Smalheiser (1997) – Arrowsmith

Since 1988, Swanson has been using computational text analysis tools based on the search strategies of his earlier work. With Smalheiser, he published several papers in medical journals detailing connections between magnesium deficiency and neurological disease (Smalheiser & Swanson, 1994), indomethacin and Alzheimer's Disease (Smalheiser & Swanson, 1996a), and estrogen and Alzheimer's Disease (Smalheiser & Swanson, 1996b).

These tools have lead to a discovery support tool called Arrowsmith (Swanson & Smalheiser, 1997). According to Swanson, Arrowsmith is not only an aid to LBD, it is a product of it, and has developed based on the search techniques and strategies described in Swanson (1988). At present, Arrowsmith includes additional enhancements from Smalheiser as well, and is available on the World-Wide Web (WWW)[6]. It is designed with capabilities that assist a user with the discovery process such as automatically extracting words and phrases that co-occur in the *A* (start literature) and *C* (targetliterature) document sets. The system is based on a three -way interaction between computer software, a bibliographic database (such as Medline), and a human operator. The interaction generates information structures that are used heuristically to guide the search for promising complementary literature structures, according to Swanson and Smalheiser (1997).

[6] http://kiwi.uchicago.edu or http://arrowsmith.psych.uic.edu

Currently, there are two different Arrowsmith systems: the first as introduced by Swanson & Smalheiser (1997), and the current version, which is open to the public via the World Wide Web (WWW). There are several improvements, and therefore differences, between the older and current versions. This section of the survey will focus primarily on the initial version, explained in detail in Swanson & Smalheiser (1997), since most of the following work is based on this system.

### Domain and Input Data

The domain is again medical literature, especially the Medline database. Arrowsmith is specifically designed for Medline with its large, domain-specific stop-word lists. Words are used as the unit of lexical analysis. Titles of the start literature (e.g., "migraine") are downloaded from Medline (Swanson & Smalheiser, 1997). Although titles are used as the main input, Medline's MeSH (Lowe & Barnett, 1994) headings and subheadings are also used in the current WWW version.

### Extracting Concepts from Text

Prior work required manual selection of physiologically significant keywords, short phrases, and meaningful word combinations from each title. This process is more automated in Arrowsmith. The extraction process still involves excluding obviously unsuitable words (Swanson, 1986), and performs the following steps:

1. Exclude unsuitable words by applying a pre-compiled exclusion list, or "stop list" to terms extracted from titles. The list is human-compiled (on the basis of judgment applied

*a priori* concerning the suitability of each word) and consists of about 5,000 words (Swanson & Smalheiser, 1997). (Note that this list is increased in size in later versions.)

2. Relative frequency is used in the selection of the intermediate concepts (*B-concepts*). Each *B-concept* becomes the basis of a Medline query, and step 1 is applied to the resulting composite document set.
3. Identifying the **terminals** (target concepts of interest) is accomplished through the number of links to intermediate topics, with topical restrictions like "dietary," "toxicity," etc.

**Approach and Model of Discovery**

The web-based version of Arrowsmith uses the closed model of discovery, even though the earlier work of Swanson and Smalheiser (1997) utilized the open model for the Migraine↔Magnesium example. This is because Arrowsmith is designed to assist in the comprehension and recognition of intermediate connections between the *A* and *C* literature. The closed approach involves two Medline searches in Arrowsmith's processing, with the first search defining the start literature (*A*) and the second defining the target literature (*C*). The program then generates a list of words and phrases found in the titles of both literatures. This list can be edited further through several steps to get *B-concepts* by:

a) selecting only certain semantic categories (e.g., anatomical regions, disorders, drugs),

b) adjusting frequency thresholds (e.g., one can select only *B-concepts* that appear in more than one paper in each literature),

c) adjusting first publication date thresholds (e.g., one can select *B-concepts* that first appeared in the *A* or *C* literature within the last two years), or

d) manual selection of concepts.

Finally, for each *B-concept* of interest, one can view the titles containing *A-concepts* and *B-concepts* ("AB titles") juxtaposed to the titles containing *B-concepts* and *C-concepts* ("BC titles").

**Evaluation**

The Migraine↔Magnesium connection is used to evaluate Arrowsmith. Since Swanson's work is a first attempt to build a LBD system, there exists no previous work to compare results or efficiency. In comparison to the original Migraine↔Magnesium study (Swanson, 1988), 10 of the 11 connections are in the generated *B-concepts* list (list of the intermediate concepts). Almost all of the originally reported *B-concept* connections were produced semi-automatically, in a matter of hours, replacing the weeks-long literature search and exploration previously required. They also observed satisfactory results in the Raynaud's↔Fish Oil connection (Swanson, 1986), but failed in the Somatomedin-C↔Arginine connection (Swanson, 1990). This emphasizes that their technique may not be always successful, even in the hands of experienced users. Although Arrowsmith is commonly used with the closed approach to investigate connections between start and target literature, the results in Swanson & Smalheiser (1997) are based on the open model of discovery experiments in which the target literature is found beginning only with the start literature set.

**Figure 5: A flowchart of closed discovery process. Adapted from Swanson & Smalheiser (1997)**

## Discussion

The Arrowsmith system provides a promising start in terms of automating the LBD process. However, there are several aspects of this work which need to be addressed in future work for establishing more general and widely-applicable systems. For one, the success of the experiments reported in Swanson & Smalheiser (1997) may be influenced by the fact that the end results were already known (the Migraine↔Magnesium and Raynaud's↔Fish Oil links). Since these tests strove to replicate the known results, the systems may have been shaped to better find these particular connections[7]. Another confounding factor is the intermittent reliance on human experts, who are also aware of the pre-known connection, and guide the tool to these results. Generalization is limited as this possibility of confounding factors or overfitting can be seen in the failure to recognize the somatomedin-C↔Arginine link. This failure may also indicate that some connections may be of a very different nature, and highlight a challenge in the development of semi-automatic or fully-automatic discovery systems.

The application scope is also limited as Arrowsmith uses only titles of the Medline records. In addition, the system in Swanson & Smalheiser (1997) uses only single words that have broad meanings, whereas multiple-word terms are more meaningful and specific. These factors, coupled with the lack of a strong lexical approach by which information is extracted from the

[7] In the data mining field, this is a form of what is termed *overfitting*.

Medline records, ranks the original system as one blazing the trail for more automated systems, but limited in terms of potential capability and effectiveness.

## 2.3. Gordon and Lindsay (1996) – Information Retrieval Techniques

Gordon and Lindsay (1996) also attempt to replicate discovery of the path leading from Raynaud's to Fish Oil with computers, both to make the discovery more automatic and extend Swanson's findings. They use techniques from Information Retrieval including token frequency (a metric summing the occurrence of tokens such as words), record frequency (the number of records (e.g., documents) containing a given token) and **tf*igf** (token frequency * inverse global record frequency, a traditional IR scaling technique). Their idea is to use these correlated statistics to identify intermediate literature with strong conceptual similarity to the starting point.

### Domain and Input Data

Gordon and Lindsay also focus on the medical domain, especially the Medline database. One difference from Swanson's prior work is that they use the complete text of the Medline records instead of just the titles. They download the set of Raynaud's documents (documents containing "Raynaud" after 1982 and before 1986) from Medline to be used as the start literature (*A*). The unit of lexical analysis is every term or two-word adjacency phrase in the start literature ("Raynaud's" in this case).

### Extracting Concepts from Text

Lexical analysis used for extracting concepts can be grouped into three steps:

1. Use of three different stop-word lists to prune the uninformative words (corresponding to “unsuitable words” in previous work): one contains the top 731 most frequently used words, a second contains 34 words with the highest number of Medline entries, and the third is a special collection of 303 “noise words” discovered in the course of repeated experimentation.

2. In the selection of the intermediate concepts, the three lexical statistics (mentioned above) for every term or two-word adjacency phrase in a downloaded literature are calculated. Those which do not meet user-selected thresholds are deemed as “noise” and removed.

3. Stemming is employed, which collapses singular and plural variants of terms. Lexical statistics are combined for terms and phrases that are collapsed by stemming. Note, however that there is no grouping of the tokens according to their meanings.

**Approach and Model of Discovery**

This model of discovery mainly follows the open model. Starting with literature about “Raynaud’s” (*A*), they search for connecting literature *B*, and from there try to reach target literature regarding Fish Oil (*C*)[8]. They identify 20-30 concepts (tokens) with the highest values for each of the statistics in the downloaded literature, as aforementioned. Then they use each of these concepts as a query to get document lists that can be used to identify intermediate

[8] As noted above, in Swanson’s first discoveries the Raynaud’s start literature was labeled as *C* and the fish oil target literature labeled as *A*. Thus the path of discovery was $C \rightarrow B \rightarrow A$. The idea is the same here but labels are used differently (i.e., reversed such that $A \rightarrow B \rightarrow C$).

literature. “Blood” was top ranked in their list, but since it is too broad they queried Medline for “blood AND Raynaud’s” and retrieved the results. The same lexical analysis was conducted in these new documents. By reviewing ranked lists for each of the three statistics, human experts suggested “blood viscosity” as an important characteristic. Finally, by looking where else “blood viscosity” is mentioned outside the Raynaud’s documents, search software helps to find “Fish Oil” as a possible treatment for Raynaud’s disease, since Fish Oil lowers high blood viscosity.

## Evaluation

They compare their results against Swanson’s Raynaud’s↔Fish Oil connection (Swanson, 1986). They evaluate performance using precision and recall on the linking concepts. Swanson’s linking concepts serve here as a gold standard. Their results confirm that Swanson’s example of Raynaud’s↔Fish Oil connection (Swanson, 1986) is discoverable, through an open process, within the medical literature.

## Discussion

In this replication study, the main difference with Arrowsmith is in the application of term frequency statistics. Gordon and Lindsay focus on identifying linking concepts based on high occurrence frequency, while Arrowsmith only eliminates terms with low frequency. This work therefore utilizes a greater set of Information Retrieval techniques in recognizing importance and linking, while Arrowsmith primarily “seeks terms whose **df** (document frequency) is significantly greater than would be predicted by a Poisson distribution” (Lindsay & Gordon, 1999).

One aspect of this work that could be clarified is the decisions leading to the acceptance of "blood viscosity" as the primary intermediate concept (*B-concept*). It is true that this phrase scored highly per the three statistics used, but there were other terms with similar performance. Plus, in their first analysis, the authors discovered the term "blood" but considered it too broad. It required a human expert to intervene and decide to search for "blood AND Raynaud's" as the following step. This seems to indicate that the system developed may not have the capabilities to modify searches appropriately, thereby limiting the user's ability to reach the desired results. This further elucidates the problems associated with building an automated system as mentioned in the Discussion in Section 2.1. The point is emphasized by the authors, however, that they tried to support discovery (not to automate it) leaving medical reasoning to those trained to do so.

A final note is the usage of Swanson's results as a gold standard to evaluate system performance. Since Swanson's results have not been rigorously developed as a "gold standard," this may not be the most effective course of action. However, with no such standards in place, this highlights the need for "gold standard" benchmarks for LBD researchers, especially as a basis for comparative evaluation.

### 2.4. Gordon and Dumais (1998) – Latent Semantic Indexing

Gordon & Dumais (1998) offer an alternative method to support literature based discovery. They use Latent Semantic Indexing (LSI), which employs latent semantics based on higher-order co-occurrence to compute document and term similarity (Kontostathis & Pottenger, 2006). LSI can reveal hidden relationships among terms, as terms semantically similar lie closer to each other in the LSI vector space. The idea is to create a Latent Semantic Index from downloaded Medline documents to examine which terms lie near (according to the cosine similarity metric) the underlying concepts (e.g., the concept of Raynaud's Disease) and draw inferences about conceptual similarity. The authors' main interest is to see if the LSI method produces considerably different or better results than Gordon & Lindsay's (1996) method of selecting intermediate literatures on the basis of token counts, record counts, and tf*igf statistics.

**Domain and Input Data**

Domain is once again the medical domain, particularly the Medline database. The authors' main interest is to compare their results with Gordon & Lindsay's (1996), and therefore they use similar input data. They downloaded the same set of Raynaud's documents from Medline for the time interval of 1983-1985. Terms and two-word adjacency phrases in downloaded literature ("Raynaud's" documents) are used as input for LSI.

**Extracting Concepts from Text**

Unlike previously discussed works, the authors make no mention of mechanisms used to eliminate obviously unsuitable words and stop words. After the LSI index is created for a particular document set (start literature *A* or intermediate literature *B*), important concepts are selected using the cosine similarity with the start concept (e.g., "Raynaud's") and a ranked list is returned to the user.

## Approach and Model of Discovery

The open model is the primary model of discovery.  Starting with literature regarding "Raynaud's" (*A*), the authors attempt to find connecting concepts which then form the intermediate literature (*B*) through querying Medline and downloading resultant records, and from there search to reach the target literature regarding "Fish Oil" (*C*), as in Gordon and Lindsay's work.

This process begins with an LSI analysis of literature *A*.  The nearest neighboring terms to the *A-concept* are identified by rank-ordering the cosine similarity of all terms to the concepts.  This ranked list is given to a medical expert who identifies intermediate concepts (*B-concepts*).  Once the *B-concepts* are selected, the documents are downloaded and this same process is applied again – this time to find the target literature.

## Evaluation

136 nearest neighboring terms to "Raynaud's", found by LSI, were compared with the union of six top-40 concept lists from Gordon & Lindsay (1996).  This was done for both one-word concepts and two-word concepts, and the size of the intersection was around 42%.  More importantly, the top-10 items in the authors' list included the first nine of the items in the list of Gordon & Lindsay (1996).  As a result, the authors conclude that there is a strong overlap among the terms uncovered by LSI and Gordon & Lindsay's (1996) method.  Overlap is strongest among the top-ranked items.  They conclude these two approaches are similar but complementary.

### Discussion

The authors evaluate their work using Gordon & Lindsay's (1996) results as a gold standard for comparison. Recall that Gordon & Lindsay's (1996) work used Swanson's results as a gold standard, which is neither fully understood nor a standard. On the other hand, the authors approach produces more precise results than others of its time. This does not solve the automation problem, however, as determining terminal concepts is still a fully manual process which must be conducted by a human expert.

## 2.5. Gordon and Lindsay (1999) – Trigrams and Contextual Analysis

Lindsay and Gordon (1999) report additional experiments that apply and extend their previous approach (Gordon & Lindsay, 1996) discussed in Section 2.3. Since the techniques they employ are similar to those in their previous paper, this section will simply provide an overview of the significant differences.

### Domain and Input Data

The same as discussed in Section 2.3, except that this current work also uses three-word phrases as a unit of lexical analysis.

### Extracting Concepts from Text

This work uses lexical analysis for concept extraction, which is based on four statistics: token frequency, document frequency, relative frequency, and **tf*idf** (term frequency * inverse document frequency). As before, stop words and noise words are excluded. Phrases including such words are also disqualified.

An important distinction from the authors' earlier work is that phrases are now always analyzed in the context of same-length phrases. This means that all one-word phrases are analyzed in the context of other one-word phrases, two-word phrases with two-word phrases, etc. The reasoning behind this change is that although a longer term (like "calcium channel blocker") is precise and descriptive, it necessarily occurs less frequently (across phrases of all lengths) than shorter phrases like "calcium," "channel," "blocker," or "calcium channel." This modification seeks to mitigate this issue.

## Approach and Model of Discovery

In this work, the authors execute two sets of experiments using the open model. They first compare their results with previous work, seeing if their methods (using the concept of "migraine") lead to any or all of the 12 intermediates linking "migraine" and "magnesium" (where 11 of them are reported in Swanson (1988) as well as vasospasm, which they consider related to vascular tone and reactivity). The second set begins analyzing each of these 12 topics to determine to what discovery targets they might lead.

## Evaluation

The authors' evaluation is similar to that of their previous work. This time they compare their results with Swanson's Migraine↔Magnesium connection (Swanson, 1988). Once again Swanson's linking concepts serve as a gold standard.

## Discussion

In this work the authors use bigrams and trigrams, which are two and three word phrases that are more meaningful than single words.  Here, the term “meaningful” conveys the sense that a bigram or trigram refers to a single, distinct and recognizable subject or item, more often than a unigram.  While bigrams and trigrams are generally more meaningful than unigrams, there are many sequences of two or three words in Medline abstracts with only a limited number of them being meaningful both linguistically and biomedically (Weeber et al., 2001).  This again dictates the use of stop-lists, which requires human effort for its creation.  Overall, the descriptive degree of longer terms and the trade off between the frequency counts may be important for future systems, and should be further studied.  By using different lengths of terms in different levels of a LBD system, the combination of their frequencies may capture more domain (human expert) knowledge from the text.

### 2.6. Weeber et al. (2001) – Using Concepts in LBD

Weeber, Vos, Klein and de Jong-van den Berg (2001) propose a two-step model of the discovery process of generating hypotheses and subsequently testing them.  In addition to this different approach, this work implements a Natural Language Processing system that uses the biomedical Unified Medical Language System (UMLS) (Lindberg, Humprey & McCray, 1993) concepts as its units of analysis.  The semantic information provided by these concepts is used as a filter. They attempt to replicate Swanson’s first two discoveries (Swanson, 1986; 1988).

### Domain and Input Data

Once again, the medical domain (particularly the Medline database) is used as the experimental environment. They use the query "Raynaud's Disease" in Medline to perform a search for comparable results from prior work, and do the same using the Migraine↔Magnesium example. Their DAD system (standing for Disease-Adverse drug reaction-Drug, or vice versa) takes raw text from the titles and abstracts and maps them to UMLS concepts using MetaMap. **MetaMap** (Aronson, 2001) is a tool that identifies biomedical concepts from free-form textual input and maps them into concepts from the UMLS Metathesaurus.

### Extracting Concepts from Text

Instead of lexical statistics, which are mostly based on document counts and token counts, this work MetaMaps titles and abstracts to UMLS concepts, and filters them through semantic filters. Co-occurrence is also used as an additional filter.

### Approach and Model of Discovery

The authors use both models (open and closed) of discovery. The main interest is in concepts that co-occur with *A* or *C* in the same sentence. The user is given a ranked list of concepts. Rank ordering of the concepts is based on concept frequency in the open discovery process. For closed discovery, the number of links between *A* and *C* are also included. No stop words or lexical statistics are used.

## Evaluation

Both Swanson's Raynaud's↔Fish Oil connection (Swanson, 1986) and his Migraine↔Magnesium connection (Swanson, 1988) are used to validate the authors' results. In the closed approach they look to see if the connecting concepts (intermediate concepts or *B-concepts*) are included in the ranked list. For the Fish Oil example, they state that the original *B-concepts* are in their list of 68 functional *B-concepts*.

In the open approach, with the Fish Oil example, they state that even though the Fish Oil concepts did not rank highly in their list, many components related to Fish Oil are found in their list with reasonable ranks. As such, a domain expert may therefore still recognize Fish Oil as a suitable target literature (*C*).

## Discussion

Gordon and Lindsay (1996) view LBD as a supporting system instead of an automated process, and the authors agree with this point of view. Their system therefore supports the human researcher in three ways: by restricting the search space, by assisting in interpretation through semantic analysis, and by providing the textual context of the hypothesis so that scientists can scan efficiently through large mounts of literature to look for new ideas or strengthen their initial hypothesis.

One of the problems with this technique, however, is that MetaMap cannot resolve ambiguous text-to-concept mappings. For example, MetaMap maps "mg" ("milligram") and "Mg" ("magnesium") both to the concept "magnesium". In comparison to Gordon & Lindsay's work, this work is more domain dependent as they limit themselves to dietary factors through the semantic filter. This involves domain-dependent expert knowledge and interest from the user.

In addition, it is worth noting that UMLS is specific to Medline and the medical domain and as a result this approach cannot be generalized to other domains which do not provide similar systems.

Finally, the authors' evaluation method did not involve any statistical or commonly applied methods like precision and recall. In their place, qualitative terms (such as "reasonable") are often used. This work would be easier to compare if more traditional and quantitative measurements were provided.

### 2.7. Gordon et al. (2001) – LBD on theWWW

Gordon, Lindsay and Fan (2001) show that LBD techniques can be used in non-traditional data sources. Using the World Wide Web (WWW) as a single large data source, they perform several experiments to display the validity and usefulness of performing LBD techniques across the Web. This broadens the applicability of LBD, as previous work almost exclusively uses the medical domain (and the Medline database in particular).

Previously discussed work displays several common characteristics in the analysis and application of LBD:

- Use of scientific literature from Medline as a source of discovery
- The discovery path is generally from disease to cure (problem to solution)

- A literature-based discovery is a connection, either completely new to the field or at least overlooked by the vast majority of its practitioners

This work also addresses these criteria, in a slightly different, yet related fashion:

- Use of scientific literature from the WWW as a source of discovery
- The discovery path is from solution to problem (finding new applications for existing solutions)
- The goal of literature-based discovery is to become more aware of connections between different bodies of information

This work therefore breaks ground in many aspects of LBD research, especially in terms of the applied domain, input data, and methodological approach.

**Domain and Input Data**

The authors apply LBD in the domain of the World Wide Web: a large not-exclusively-scientific literature source. In their first experiment, they locate and download the 50 most prominent documents on the WWW related to the topic "Genetic Algorithms," as returned through the AltaVista[9] search engine. The second step of the authors' approach repeats the Web search but substitutes four topics known to be central to the idea of genetic algorithms in place of the term "Genetic Algorithms." The raw data for the LBD system is the textual content of the retrieved web pages. From these pages, two-word phrases (bigrams) are parsed and utilized as the base units of analysis.

[9] The AltaVista engine is located at http://www.altavista.com, at the time of writing.

## Extracting Concepts from Text

Through their lexical analysis, the authors obtain two basic statistics: **token counts** (number of times a phrase occurs within a single web page) and **document counts** (number of retrieved web pages containing a given phrase). These are used to derive several other statistics, which are applied in forming a ranked list of phrases. Prominent concepts are then selected by a human expert. This signifies that human judgment is again used in this part of the process, similar to the previous works.

## Approach and Model of Discovery

Their approach is to take a subject, often a technology or technique, then use LBD techniques to analyze the selected domain and attempt to see how it might be applied to problems in new areas.

This work employs the open model of discovery, similar to that of most previous work ($A \rightarrow B \rightarrow C$). Beginning by selecting 12 prominent concepts related to genetic algorithms (start literature *A*), statistics are generated for each concept as outlined in the "Extracting Concepts" section above. These concepts are applied as queries to AltaVista, and the top 100 resulting web pages returned were stored. Pages are parsed and lexical statistical analyses produced. The authors pool these 12 lists of analyzed web pages using normalized token frequency. A human expert then selects 42 prominent concepts from the set. Each of these concepts can be a potential discovery if they are not mentioned together with the start concept. The intersection between these concepts and the start literature is determined by using the Web of Science (1987 to present) and UseNet to search for co-occurrences of each concept and the term "genetic algorithms". In their experiment, several non-intersecting items were identified. Ultimately, a target concept may be any of the non-intersecting concepts.

### Evaluation

While this work pushes the boundaries of LBD research and application, very little information is provided concerning evaluation of the results or technique. This may be due to several reasons, including the usage of a new domain, new intermediate statistics, and a new end-goal for LBD. Therefore, there is no similar work available for comparison. Anecdotally, the authors do ask John Holland, a pioneer in genetic algorithms research, to identify items that may be interesting to study in conjunction with genetic algorithms. Compared to the final results of one of their experiments, however, none of the items identified by Holland appears on their list.

### Discussion

Overall, this work provides the extension of LBD outside of the medical domain, to the World Wide Web. However, there is no evaluation of the accuracy, usefulness, or correctness of the results. Also missing is an evaluation of the technique applied, as it is neither compared nor contrasted with others. Finally, as a human expert is still required for the selection of intermediate prominent concepts, this does not provide a method for complete automation.

## 2.8. Pratt and Yetisgen-Yildiz (2003) – LitLinker

Pratt and Yetisgen-Yildiz (2003) present an LBD system, LitLinker, which incorporates knowledge-based methodologies, natural-language processing (NLP) techniques and a data mining algorithm. This is one of the few LBD studies that employs a data mining algorithm. Specifically, they use association rule mining, or ARM (Agrawal, Manilla, Srikant, Toivonen & Verkamo, 1995)[10]. Interestingly, association rule mining is an unsupervised learning technique very similar to co-occurrence analysis, the primary difference being that in ARM the tri-occurrence, quad-occurrence, etc. of terms is discovered. Thus this data mining algorithm is used to find correlations between concepts. The authors' knowledge based methodologies use Medline's knowledge base, the Unified Medical Language System (UMLS).

**Figure 6: The Text Mining Process in LitLinker (Pratt & Yetisgen-Yildiz, 2003)**

## Domain and Input Data

This work also utilizes the medical domain and the Medline database in particular. Medline is queried with the initial concept ("Migraine", in this case), and only the titles of the returned documents are stored and used in further analysis. This is a significant distinguishing factor as the amount of content is substantially reduced when the full Medline record is ignored in this way.

## Extracting Concepts from Text

Extracting tokens from the titles of the Medline documents is accomplished using UMLS, similar to previous work (Weeber et al., 2001). LitLinker also uses the MetaMap tool to map free text to biomedical concepts, also like the previous work, which is designated as the "Parse Titles" action in Figure 6. The process again leverages the UMLS hierarchy, using it to prune concepts that are

[10] ARM was applied to LBD previously by Hristovski, Stare, Peterlin and Dzeroski (2001).

too general. The authors observe that many general concepts, such as "disease" and "drug" appear on the second level of the UMLS hierarchy as children of root concepts (which are biomedical vocabularies) as well as on the third level (children of the concepts on the second level).

After this first pruning operation, many general terms remain, so additional pruning is performed. The second operation prunes concepts that appear in the titles of more than 10,000 MEDLINE documents. Also, concepts that are too closely related to the start term are eliminated by using UMLS to determine all the parents and children of the start concept. Third, concepts that make implausible connections are pruned by using UMLS semantic types.

Once all pruning is complete, similar terms are grouped together to raise the frequency of the central concept (which is fully realized across many words and phrases) to a more significant, noticeable level. They cluster the related concepts into a group labeled with the shortest concept name. The result of these operations is a set of tokens to investigate correlations. As can be seen from the above processes, UMLS is a central component to this process.

**Approach and Model of Discovery**

The authors' approach follows Swanson's classic path of discovery of $A \rightarrow B \rightarrow C$. This process distinguishes itself from others in that it uses a data-mining algorithm to identify correlations among concepts, and then uses these correlations for open-ended discovery. The authors use the open approach as their model of discovery. They also use knowledge-based methodologies (such as pruning and clustering related concepts using UMLS) to limit the search space.

**Figure 7: Discovery Process in LitLinker (Pratt & Yetisgen-Yildiz, 2003)**

**Apriori**, a popular association rule mining algorithm, is used in this work. The authors set the minimum support level to 0.002, which translates to a concept occurring in at least five titles, in this context. Beginning with a given start concept, any correlated concepts found in this step act as the linking concepts. The same process is repeated using each linking concept for separate literature searches. The resulting correlated concepts for each linking concept create the total set of target concepts.

### Evaluation

Swanson's Migraine↔Magnesium example (Swanson, 1988) is used to evaluate the authors' results. The target concept, "Magnesium", is reached at a rank of 11 in their list, using an open model of discovery. Rank is computed by the number of concepts that link them to "Migraine", the start concept. In a closed model experiment, LitLinker is able to identify five of the eleven connections in Swanson's earlier work.

### Further Work

In Skeels et al. (2005), the authors extend their previous work and discuss the importance of graphical user interface and other design issues on LBD systems, as well as the development of the interface for LitLinker. Their design is based on several user-oriented goals, including promoting user comprehension of the complex relationships amongst the terms involved in each proposed connection, flexible navigation, and different levels of detail for evaluation of the connections. They also conducted a usability evaluation. Utilizing the results of this evaluation, they redesigned their interface and developed a web-accessible version.

**Discussion**

Unlike previous systems that explored the Migraine↔Magnesium example, the authors' system is able to find the connection in a more automated open-ended discovery process. Their method increases the degree of automation because human intervention is not needed in selecting *B-concepts*. Although their system is more automated than previous systems, they employ the same classic *A*→*B*→*C* approach of Swanson in discovery. Also similar to previous work, the authors employ no robust methodology for evaluation, such as those based on ten-fold cross validation, precision, recall or the traditional support and confidence metrics employed in association rule mining. The possibility of overfitting thus arises in this context as well.

**2.9. Srinivasan (2004) – Generating Hypotheses from Medline**

Srinivasan (2004) follows the discovery framework of Swanson and Smalheiser (1997). A key difference is that his algorithm relies squarely on Medical Subject Headings (MeSH) (Lowe & Barnett, 1994) (considered as the metadata of the Medline records) and UMLS semantic types. Most of the previous work uses free-text portions of Medline records with some using MeSH in a secondary role, at best. The author's motivation is to determine the effectiveness of an approach that almost completely relies on MeSH, thereby reducing the amount of manual effort involved during the discovery process. The goal of the author's system is to automatically return a ranked list of concepts to the user with potentially interesting concepts appearing at the top of the list.

**Figure 8: Semantic types of MeSH terms (Srinivasan, 2004).**

### Domain and Input Data

This work also falls within the medical domain, via Medline. However, instead of Medline titles, abstracts or full records, they use only the metadata portion of the Medline record, referred to as the Medical Subject Headings (MeSH). MeSH terms are assigned to the records by trained indexers at the National Library of Medicine (NLM) and therefore are assumed more likely to be biomedically relevant and accurate.

### Extracting Concepts from Text

The important concepts are taken directly from the terms in MeSH, and do not require a special process of extraction from narrative text. MeSH is a detailed, all encompassing system, with the MeSH vocabulary being classified using 134 UMLS semantic types such as Cell Function, Sign or Symptom, etc. Each MeSH term is assigned one or more semantic types: e.g., “interferon type II” falls within both “Immunologic Factor” and “Pharmacologic Substance.” These MeSH terms therefore provide specific semantic and relational information, which is directly used in the analysis process.

### Approach and Model of Discovery

Topics are built from MeSH-based profiles. Profiles are simply vectors of weighted MeSH terms. The author separates MeSH terms by semantic type and computes term weights within the context of a given semantic type. He uses tf*idf (term frequency * inverse document frequency) for the weighting scheme. A profile therefore represents the relative importance of different MeSH terms, within given semantic types, associated with a topic's document set.

This work performs experiments on two open and five closed discovery problems. The user of the system first identifies desired semantic types that are then used as a filter to reduce the search space. In the open discovery model the author follows the *A*→*B*→*C* type of discovery and directly (without human intervention in selecting *B-concepts*) provides a ranked list of MeSH terms to the user. The resulting items are scored by the number of the paths connecting the target concepts (*C-concepts*) back to the start concept (*A-concept*), and the strength of those paths. This scoring is used to rank *C-concepts* within each semantic type by their total combined path weights.

In the closed discovery approach, profiles are built for *A* and *C* that are constrained by the semantic types for *B*. In so doing so, terms belonging to given user-specified semantic types may potentially link *A* and *C*.

**Evaluation**

This work claims that the use of only the MeSH metadata field of Medline is competitive with the free-text based methods explored by others. The author supports this claim by demonstrating the ability to rank the key MeSH terms within the top 10 terms for both of the open discovery experiments conducted. The first open discovery experiment involved finding the Fish Oil↔Raynaud's Disease connection (Swanson, 1986). A semi-automatic open procedure was applied (as in Gordon & Lindsay, 1996 and Weeber et al., 2001), followed by a closed approach as in Weeber et al. (2001). The author also evaluated his system with the Migraine↔Magnesium example using both open and closed discovery approaches. In addition, the author evaluated a closed approach for the Indomethacin↔Alzheimer link (Smalheiser & Swanson, 1996a), the Somatomedin C↔Arginine connection (Swanson, 1990), and the Schizophrenia↔Calcium-Independent Phospholipase A2 link (Smalheiser & Swanson, 1998a). In open discovery, as noted the author's approach successfully ranked key MeSH terms in the top 10 positions within the appropriate semantic types for both experiments (Raynaud's↔Fish Oil and Migraine↔Magnesium). In the closed model, after semantic type filtering, most of the terms ranked within the top 10 of a particular semantic type of a user-selected *B-concept*. The author points out the difficulty of comparing his results with previous work because of the difference in ranking strategies. He defines normalized term weights within each semantic type and these types define independent term groups, while other researchers propose a single ranking of all terms regardless of semantic grouping. The author nonetheless claims that, in general, his term rankings are consistently better than those in previous results.

**Further Work**

In follow-on work in Srinivasan & Libbus (2004), the authors apply their LBD algorithms to determine the effectiveness of their open approach in revealing novel connections that might lead to reasonable hypotheses. Similar to their previous work, the authors' algorithm follows the discovery framework first developed by Swanson and Smalheiser (1997). In their current study, the authors start with *curcuma longa*, a dietary substance and widely used spice in Asia, to explore therapeutic potentials. In doing so, they identify several diseases and disorders that could be used in forming new and testable hypotheses.

## Discussion

This work, in deviating from most prior work and utilizing only metadata of a domain, displays fairly impressive results. By consistently providing results in the top-10 items of their lists, they demonstrate that a metadata-only approach is feasible and effective. Nonetheless, as is the case in previous work, the evaluation is hampered by the lack of usage of gold standards, and established metrics and methods of evaluation. In addition, relying solely on this higher level of metadata leads to two major points of concern and discussion. First, not all literature sources possess such metadata, or at least not at the same level of depth and specificity as UMLS. For example, if applied to the WWW, this approach might not be feasible as web page metadata (as given in the <meta> tags) is free, unstructured text, and more structured formats (such as the Semantic Web (Berners-Lee, Hendler & Lassila, 2001)) are still under development. Second, MeSH terms are sometimes more general than the actual concepts that appear in narrative text. Differences in granularity may be an important factor in the use of LBD systems, and it is possible that granularity information can be useful for future systems.

## 2.10. Van der Eijk et al. (2004) – Associative Concept Spaces

Van der Eijk et al. (2004) propose a novel algorithm for finding associations between related concepts present in literature. This work is very different from other works previously discussed for several reasons: the user output is a visual graph displaying closeness of concepts instead of a ranked list, the classic $A \rightarrow B \rightarrow C$ approach is not directly followed, and concepts are analyzed using a multi-dimensional algorithm. In particular, concepts are mapped to a multi-dimensional space by a Hebbian learning algorithm using co-occurrence data as input.

Co-occurrence is a central concept to this approach. The assumption is that related concepts co-occur more frequently in articles than non-related concepts. As such, much of the authors' analysis involves searching for paths between concepts. A path is defined as a chain of co-occurring concepts. Paths that connect concepts in more than one step are indicative of an indirect relationship between two concepts. The authors claim that by exploring co-occurring concepts and searching for paths between concepts, scientists may find indirectly related scientific papers.

### Domain and Input Data

The domain is once again the medical domain, specifically Medline as in many of the previous works. Abstracts of the Medline records are used as input for analysis. The authors' indexing algorithm first extracts sentences and then eliminates stop-words. Remaining words are normalized by reducing nouns to the singular and verbs to the first person singular form.

### Extracting Concepts from Text

Normalized terms or phrases are identified using the Medical Subject Headings (MeSH) 2002 thesaurus. A unique concept identifier from MeSH is assigned to each term or phrase.

### Approach and Model of Discovery

The authors employ compact representations of documents termed **fingerprints**. The unique concept identifiers associated with the terms in each document are included in the fingerprints. Each concept identifier is assigned a relative score based on the term frequency and specificity of the term in the thesaurus (depth in the hierarchy). Once all documents are fingerprinted and all unique identifiers included, the co-occurrence of concepts in fingerprints is used to form an Associative Concept Space (ACS).

### Visualization

The ultimate end-product of this approach is the visualization of concept relations in a prominent and meaningful way. This is a non-trivial problem, as any naïve or random placement of concepts would make relative positioning completely arbitrary. A method or model is therefore required to provide meaning and purpose to these relative positions, as this positioning may reveal important information when exploring literature for novel relationships. The authors describe a mapping from a co-occurrence graph to an ACS, in which concepts are assigned to a position in space in such a way that the stronger the relationship between concepts, the closer they lie in the ACS. In other words, concepts that are connected by several co-occurrence paths, either directly or indirectly, have a small distance in the ACS. Determining appropriate positions for these concepts is accomplished through a Hebbian learning algorithm. One result of this approach is that two concepts which never appear in the same article, yet have many co-occurring concepts, should be very closely placed in the ACS, thereby allowing the visual inference of interesting relationships. Examples of outputs can be seen in the following figure.

**Figure 9: Two-dimensional projection of part of an 8-dimensional ACS from a set of Medline abstracts on Muscular Dystropy. The ACS suggests a relationship between "Insulin" and "Ferritin" (Van der Eijk et al., 2004).**

## Evaluation

The performance of the authors' algorithm is evaluated using simulated data sets for which the outcome is predictable. This artificial test data is generated using a model of scientific literature (reflecting a pattern which they observed in their Medline experiments). This model consists of 10 clusters, with each cluster representing a field of interest, characterized by a set of specific concepts. Forming an ACS with this dataset should provide a resulting space in which concepts from the same cluster are separated from other concepts. The separation of the clusters is achieved through an invariant scattering criterion from cluster analysis (Duda & Hart, 1973). When the mapping is applied to this dataset, concepts from different clusters are well separated after 20 cycles. Also, two experiments are conducted on real data from Medline to evaluate the algorithm with actual literature. Instead of using previous well-known examples of LBD works such as Raynoud's↔Fish Oil or Migraine↔Magnesium, they query Medline with "Duchenne OR DMD OR dystrophy OR limb-girdle OR LGMD OR BMD". From the visual output, the authors conclude that the ACS algorithm reveals implicit associations between medical concepts, which are explicit in several Medline abstracts but only implicitly present in the subset they use. Figure 9 provides an example of this conclusion.

**Discussion**

Although visualization is common in many areas of computer applications – even commercial products (e.g. ClearForest[11] which uses visualization to depict higher-order links), this is the first work to be reviewed that employs a visual technique in LBD. This paper does not follow the classic Swanson approach of *A*→*B*→*C*, but rather searches for chains of co-occurrence. A trivial example of this is if textX and textY are found in record1, and if textY and textZ are found in record2, then there is a chain of co-occurrence from textX to textZ. They also mention that evaluation on real data, like existing scientific documents, is complicated because it is extremely difficult to establish a reference for the explicit and (particularly) the implicit associations present in a set of documents.

The evaluation method employed is distinctfrom previous work but could nonetheless have followed prior results. For example, an attempt to visualize the Migraine↔ Magnesium relationship could have been performed and then used to compare the effectiveness of this work with the previous work in terms of finding implicit connections. This path has not been followed however, and again the overall evaluation methodology suffers from lack of use of standardized techniques.

### 2.11. Wren et al. (2004) – Random Models

One of the more recent applications of the LBD techniques (Wren et al., 2004) finds a connection between "cardiac hypertrophy" and a chemical compound named Chlorpromazine

[11] http://www.clearforest.com/

(CPZ). That connection suggests CPZ (a commonly used drug) has an unrecognized effect (anti-hypertrophic in nature) on cardiac hypertrophy. This relationship is novel and has not been previously suggested in the literature. The authors pursue this finding by searching for associations between these two in a clinical experiment (using rodents) in order to support this hypothesis.

### Domain and Input Data

This work also falls within the medical domain, again using the Medline system. A significant difference from prior work is that all electronically available Medline records are used in the analysis, as opposed to separate literature sets (e.g., sets of documents downloaded by querying for *A*-, *B*-, or *C-concepts* in Medline). The authors use the free-text of both titles and abstracts of the Medline records for concept extraction.

### Extracting Concepts from Text

Concepts are extracted (e.g., genes, diseases, phenotypes, chemicals) by mapping free-text to concepts via several reference databases including OMIM (Hamosh et al., 2000), MeSH (Lowe & Barnett, 1994), LocusLink (Maglott et al., 2000) and HGNC (Povey et al., 2001). Acronyms are also resolved using an Acronym Resolving General Heuristic (Wren & Garner, 2002).

### Approach and Model of Discovery

Similar to the Swanson's open approach, the authors start with *A-concepts*, and identify *B-concepts* by identifying the co-occurrence of these concepts within Medline records. Each *B-concept* is used as a query to identify *C-concepts* by using co-occurrence metrics again. These *C-concepts* are related to the *A-concept* only implicitly, in that any relationships with *A* are not currently documented in the literature. There may be a very large list of *C-concepts* implicitly related to *A*, and each may be a potential discovery of a new relationship. These relationships are ranked by comparing the discovered *A-C* relationships against a random network model. This is accomplished by dividing the number of observed connections (in Medline) by the number of connections expected by chance, as demonstrated through the random model. Those relationships whose quotients surpass a threshold of statistical significance are ranked more highly in the final list. In this experiment, the authors begin with the *A-concept* of "cardiac hypertrophy" and return 20 concepts with the most implicit connections, sorted by the observed to expected ratio. The *C-concept* "Chlorpromazine" is ranked third in this list.

The authors claim that this ranking correlates with the probability that two objects are related and with the strength (frequency of co-occurrence) of that relationship. The probability of significance of a co-occurrence is defined as: P(related) = 1- $r^n$ where *n* is number of co-occurrences, and *r* is the error rate that is estimated by manually evaluating the co-occurring concepts within a random set of 25 Medline records. They also define "strength" as a function of the number of times two concepts co-occurred and the aforementioned probability.

**Evaluation**

The authors evaluate their results by performing a clinical study on mice. The authors do not discuss why they return only 20 items in the resultant list, nor why they choose not to consider the terms ranked below "Chlorpromazine" (ranked 3rd) for further study. Along with these unexplored analysis choices, this work also suffers from the absence of standard evaluation methodologies, similar to the previous work.

## Discussion

The authors' technique attempts to discover statistically significant connections between *A* and *C* by analyzing shared connections between *A-to-B-concepts* and *B-to-C-concepts*. One advantage of their method is the functionality for resolving ambiguous acronyms, thereby affecting the number of co-occurrences between concepts and providing what is perhaps a better reflection of the importance of the concepts. They also leverage several databases to extract only medically meaningful concepts from titles and abstracts of the Medline records, thus reducing the problem space. Their method automates the LBD process by identifying implicit relationships (e.g., *A*↔*C*), thereby removing the need for human intervention when selecting appropriate *B-concepts*. As noted, however, similar to other approaches, a robust methodology for evaluation is lacking.

## 2.12. Hristovski et al. (2005)– BITOLA

Hristovski et al. (2005) present BITOLA, an interactive literature-based biomedical discovery support system. Building on their previous work (Hristovski et al., 2001), the authors integrate genetic knowledge about the chromosomal location of the starting disease as well as the chromosomal location of the candidate genes to enable their system to perform disease candidate gene discovery.

### Domain and Input Data

This work is in the medical domain (including Medline), with an additional focus on the genetic literature. The authors used MeSH descriptors, titles, and abstracts of the Medline records. Although the main data source is Medline, LocusLink and Human Genome Organization (HUGO) are used to aid in the appropriate mapping of gene names and chromosomal locations.

### Extracting Concepts from Text

Gene symbols are extracted from the title and abstract of Medline records, and form the input concepts of the system with the addition of appropriate MeSH descriptors.

### Approach and Model of Discovery

The authors used association rules between pairs of biomedical concepts to discover known and unknown relationships. Following the classic flow of Swanson's *A*→*B*→*C* (open discovery approach), the algorithm used in the BITOLA system is as follows (adapted from Hristovski et al., (2005)):

1. Let *A* be a given starting concept of interest

2. Find all concepts, *B***,** such that there is an association rule *A* → *B*
3. Find all concepts, *C***,** such that there is an association rule *B* → *C*
4. Eliminate those*C-concepts* whose chromosomal location does not match the location of the starting concept *A*
5. Eliminate those*C-concepts* for which an association *A* → *C* already exists
6. The remaining *C-concepts* are candidates for a new relation between *A* and *C*
7. Rank and display the remaining *C-concepts*

Ranking the remaining *C-concepts* is accomplished through the use of the common support metric for association rules (Agrawal et al., 1995). The authors also use filtering, incorporating semantic types and applying thresholds based on the support and confidence metrics for association rules. This is needed in order to limit the number of *A*→*B* or *B*→*C* associations since the number of such rules can be extremely large.

Since their system includes gene information, the authors' process can be applied in several promising scenarios. First, a user of the system may start with a genetic disease for which the global chromosomal region is known (but not the exact gene location), and by linking a *B-concept* (e.g., pathological or cell function), the user may find a gene within the same region or expression location. Also, a user may start with a known gene and search for diseases that may be caused or affected by that gene, using a similar *B-concept* as above.

Finally, the authors propose a method to help disambiguate gene symbols in a large number of Medline documents.

### Evaluation

No evaluation is provided in this paper, but future evaluation is planned. This evaluation plan first requires identifying recently discovered disease-gene relationships. Next, the authors plan to determine if their system is able to detect such relationships, given only the medical information known (the documents published) prior to the time of discovery. The database of disease-gene relationships is a ground truth in this case, and as before due to potentially limited examples of the concept to be learned, there is a possibility of overfitting unless standard techniques such as cross-validation are employed.

### Discussion

The authors indicate that although BITOLA is a LBD tool in general, it is particularly useful for finding new relationships between diseases and genes because of the integration of gene background knowledge (e.g., chromosomal locations). They also use the complete Medline database, which consisted of about 11 million records at the time of their publication.

## 3. Trends in Literature Based Discovery

In this section we will observe some trends that cross multiple works surveyed. First, Swanson’s pioneering work provides the framework on which almost all work in LBD is based. Swanson’s initial discoveries were made through a laborious, time-intensive, manual process. Further work endeavored to mitigate these challenges by developing processes to make LBD easier to perform, faster, and overall more automatic. In doing so, different techniques for concept extraction, computation of results, and sizes and types of input data were utilized. Initially, the role of the human expert remained significant, and the systems followed Swanson's *A*→*B*→*C* model of discovery. Still, incremental improvements occurred as each system contributed a technique, approach, or metric designed to ease or automate the LBD process.

One trend that is developing is the use of two or more words per concept instead of just one. In their earlier work, Weeber et al. (2001) for example use single-word terms rather than multiple-word terms such as bigrams or trigrams. Later efforts however increasingly employ multiple-word terms. A related trend is in the increasing use of specialized knowledge bases that are part of the UMLS. This provides more informative and useful terms than those extracted based on simple frequency statistics.

More recent work has focused specifically on, and provided advancements in, automation of the LBD process. Using more advanced Natural Language Processing (NLP) techniques while at the same time exploiting metadata (e.g., from UMLS) has led to a reduction in the role of the human expert. One of the most recent works, Srinivasan (2004), relies completely on UMLS semantic types and the Medical Subject Headings (MeSH), both used as metadata for Medline records. Similarly, Van der Eijk et al. (2004) identify concepts using a MeSH thesaurus, which is a part of the UMLS system. This system also partially automates the discovery process using a mapping from a co-occurrence graph to an Associative Concept Space (ACS). Other recent work in which the role of the human expert is reduced include Pratt & Yetisgen-Yildiz (2003), Srinivasan (2004), Wren et al. (2004) and Hristovski et al. (2005). In each of these approaches the system generates a ranked list of candidate discoveries without human intervention in selecting appropriate intermediate concepts (*B-concepts*).

A related trend is to use more advanced methods to capture important correlations between concepts. Hristovski et al. (2001), Pratt & Yetisgen-Yildiz (2003), and Hristovski et al. (2005) used an unsupervised machine learning algorithm (association rule mining) along with (in some cases) support and confidence metrics. In contrast, Wren et al. (2004) used statistical techniques to distinguish significant correlations. A related trend is the application of visualization. Van der Eijk et al. (2004) for example differs from other work by giving a visual output directly to the user without the intermediate steps requiring human expert guidance. Overall, reducing reliance on human experts by increasing the degree of automation is an important recent trend in LBD research.

These trends reveal three areas where further research is needed in LBD. First, a better theoretical understanding of LBD is needed. Swanson's pioneering work demonstrates the potential and provides a methodology by which to perform the process, but little has been accomplished in terms of evaluating theoretical foundations for the process – e.g., determining a theoretical basis for utilizing higher-order co-occurrences in LBD. Second, there is a critical need for evaluation methodologies that employ appropriate metrics and gold standards. Additional theoretical work is required to address this issue as well. Third, further automation of the LBD process is generally accepted as the proper course of action. While not all agree that complete automation should be the ultimate end-goal for LBD, the ability to provide automation could provide faster processing times and result in the establishment of a larger knowledge base on which to base further work. In what follows, we discuss these three areas in more depth, including suggestions for future work.

## 4. Theoretical Foundations for LBD

Little if any underlying theory has been developed for understanding LBD. As is clear from this survey of the state of the art, the research is primarily applied in nature, and as a result does not contribute much to a theoretical understanding of LBD. This trend in LBD research can be partially understood by considering the difficulties facing LBD researchers in developing theoretical foundations.

First, there is little (if any) detailed knowledge about the nature of the discovered connections. This leads to several questions, such as: how many and what kind of concepts are connecting *A* and *C*? What are strengths of these connections? How descriptive are the linking concepts (*B-concept*s)? Second, and perhaps more importantly, is there a common pattern in these discoveries? If so, is it general enough to apply to other domains, systems, or processes?

As pointed out in Gordon & Lindsay (1996), there are numerous associated "intermediate" literatures. Examining any of these may reveal many others. Figure 10 provides an example of this situation although only two "paths" are shown. Two questions predominate that are common to searching in general: "Which path should be followed?" and "How far should the path to be explored?"

**Figure 10: Multiple search paths with multiple intermediate literatures. Adapted from (Gordon & Lindsay, 1996). Examining many *B-concepts* may reveal many others and carry us to different targets.**

Swanson's *A*→*B*→*C* approach to discovery may not be the only way to reach the target(s). There may be different pathways, or more levels between *A* and *C*. Swanson's approach is based on second order co-occurrence, which is defined as follows (in the context of LBD): an *A-concept* (start concept) co-occurs with a *B-concept* (intermediate concepts) in one or more documents (within the *A* literature). *B-concepts* co-occur with *C-concepts* (in the *B* literature). Yet, to be a novel discovery, *A-concept* and *C-concept* must not co-occur anywhere in the entire literature (e.g., Medline).

Higher order co-occurrences may also be used to discover novel connections. Figure 11 depicts first, second and third-order co-occurrence. Such higher order co-occurrences are used in other fields, including Latent Semantic Indexing (LSI), which is a text indexing technique pioneered by Deerwester et al. (1990) and first used in LBD by Gordon & Dumais (1998). In our previous work in Kontostathis & Pottenger (2006) we prove that LSI implicitly depends on higher-order co-occurrences and show that higher-order co-occurrences play a key role in the effectiveness of systems based on LSI. As can be seen from the LBD work than employs LSI (Gordon & Dumais, 1998, in Section 2.4), orders of co-occurrence greater than two can (as one example) underlie second order co-occurrences in Swanson's *A*→*B*→*C* framework. In addition, as noted above higher orders of co-occurrence can connect start and target concepts in LBD – orders higher than two. This then raises several interesting theoretical questions: "How many higher-order paths link start and target concepts? What is the nature of the concepts in such links? What are the strengths of the connections (and how can strength be measured)? Are there any common higher-order patterns in the linkage?" Our ultimate research goal is to uncover the theoretical foundations for LBD-based knowledge discovery within a framework of higher-order co-occurrences. Based on our prior work dealing with the theoretical foundations for LSI in Kontostathis & Pottenger (2006) as well as in supervised knowledge discovery, we believe we are positioned to make significant progress towards this goal.

**Figure 11: Higher order co-occurrences (from Kontostathis & Pottenger, 2006)**

## 5. Evaluation Methods and Gold Standards

Evaluation of systems, such as the LBD process, is a multifaceted task. Knowledge-based systems pose an additional fundamental challenge in evaluation because, if they are successful, then by definition they are capturing new knowledge that has yet to be proven useful (Pratt & Yetisgen-Yildiz, 2003). Evidence supporting the preliminary discoveries of Swanson was provided later by medical researchers – the initial discoveries were made by trial and error (Gordon & Lindsay, 1996). Providing evidence in support of such discoveries is only one perspective of evaluation. Evaluation can also be based on the generated results. The correctness of the results returned, often measured as accuracy and/or precision, is one such metric. Recall is another, which refers to the number of correct results returned compared to the total number of correct results available. Other metrics reflect more qualitative aspects of the system, such as complexity of the user interface. For systems whose aim is to support human experts in the discovery process, usability issues are very important. Finally, the degree of automation in reducing the human expert's role in LBD systems is an important evaluation criterion for the effectiveness of much work that follows Swanson's pioneering research.

In surveying the LBD research that followed Swanson's work, it is clear that most approaches rely on a comparison of their own results with those of Swanson.  This restriction in the majority of the aforementioned research provides a common (albeit limited) baseline framework for comparison across systems.  Even with this framework, however, evaluation of these systems is not straightforward and important fundamental questions arise: "Are Swanson's first discoveries an appropriate 'gold standard' for LBD research?  Most of the aforementioned work uses these connections for comparison and evaluation, but is this appropriate?"  **Gold standards** are ground truth datasets that capture the correct classification of items.  In the case of LBD, a gold standard would contain pairs of start/target concepts that have been demonstrated to be novel within a given timeframe.  While evidence has been collected supporting Swanson's first discoveries, there is no known evidence suggesting that these connections, discovered by trial and error, are the only novel links connecting these literatures.  Plus, as newer and more automated processes are developed, it is uncertain if the newly discovered connections are "better," "worse," or "equivalent" to Swanson's findings.  In fact, no metrics exist to define "better" or "worse" in this context!  In other words, there is no prevalent ranking scheme to determine the overall goodness of an LBD system across algorithms and domains.

These difficulties stem from a dearth of research into the theoretical foundations for evaluation of LBD systems.  Although empirical evidence can be gathered in support of a novel connection, the evaluation methodologies employed in clinical studies do not necessarily translate to the domain of text mining. Further theoretical understanding of both LBD and its evaluation will provide a basis upon which to effectively and accurately compare various approaches to LBD. Continued study of this nature will lead to the development of appropriate gold standards and effective metrics and methodologies for evaluation. In this way progress can be made in the field, including the development of more automated systems.

## 6. Discussion of Automation of the LBD Process

The concept of LBD is easily applicable to many domains. In fact, LBD can be used for almost any kind of discovery in any domain. One of the major issues in the widespread adoption and execution of LBD, however, is automation. In its early form LBD was a laborious process in terms of the time, energy and manpower required to make even a single discovery. Later work has come far, placing a significant fraction of the computational burden on the computer, while requiring humans to provide input only at key decision-making points (such as "which path to follow" or, in later work, adjusting filters and selecting appropriate *C-concepts*). Still, validated LBD systems that automatically extract appropriate data, make appropriate decisions and display appropriate results remain elusive.

There are many domain, data, and goal-specific challenges in fully automating LBD systems. As such, for this discussion, the scope of these challenges will be limited to the Disease↔Characteristics and Symptoms↔Cure-or-treatment examples. In this classic discovery framework, characteristics, symptoms and causes of diseases must first be extracted from the literature. Because these types of entities appear primarily in free text, this is an information extraction process. Because important characteristics should be often mentioned in a particular literature set, these concepts' frequencies need to be high. Much of the work surveyed uses statistics to extract important concepts, but other issues and methods exist. One example: for a given concept, how important is its level of precision to the rest of the process? In building any system, extracting descriptive concepts from start literature should be a high-priority first step. There are several examples in the information extraction literature for extracting descriptive concepts such as key-phrases from free text documents (Frank et al., 1999). In Wu et al. (2003), important characteristics of solutions to various problems are extracted from patent data using a Reduced Regular Expression Discovery algorithm. This same supervised learning algorithm can be used to extract important concepts from free text in scientific or medical literature. After extracting descriptive concepts (e.g., a disease in LBD), the conceptual space can be searched for second or higher order connections to other concepts. A plethora of other approaches to information extraction exist, and any sensible attempt to automate LBD must leverage such technology.

As in the construction of any model using supervised machine learning, properly labeled training data must be acquired. In the case of LBD, focus on a particular domain may improve classification accuracy. Currently, six true positive instances are available in the medical domain: Raynaud's↔Fish Oil, Migraine↔Magnesium, Indomethacin↔Alzheimer's disease, Somatomedin C↔Arginine, Schizophrenia↔Calcium-independent PhospholipaseA2, and Therapeutic Uses for Thalidomide↔Four Diseases. Note that, in the final example, Interleukin-12,a *B-concept*, was selected by domain experts as an important immunologic pathway that may result in new applications for thalidomide (Weeber et al., 2003). Clearly the lack of extensive, labeled training data has inhibited progress in the application of supervised learning algorithms in LBD. In fact, this is one of the primary issues that must be overcome if progress is to be made in automating LBD using such algorithms. This point hearkens back to the identical point made earlier concerning the lack of substantial ground truth datasets. In the context of supervised machine learning, a **ground truth** dataset is a set of labeled instances representative of the concept to be learned. Each **instance** is composed of a set of attribute values combined with a class value, which in this case is a simple 'yes' or 'no', denoting that a given instance is or is not a novel discovery. The issues go further than just the need to properly label instances in the training data, however. Based on our survey, the unasked (and certainly unanswered) question is "What are the attributes of a novel discovery that are highly correlated with the 'yes' class?" In other words, the question is which attributes must be selected. As outlined in Witten & Frank (2000), this process itself can be partially automated using machine learning algorithms for attribute subset selection. Nonetheless, even this process is often facilitated by domain expert selection of an initial set of relevant attributes.

Thus the key issue revolves around the selection of attributes correlated to the class. In partial answer to this question, deeper co-occurrence analysis may provide a first step in selecting appropriate attributes. In the previous section we defined second, third and higher orders of co-occurrence. In our previous work in Kontostathis & Pottenger (2006), we determined that second order co-occurrences play an important role in the performance of LSI. The question that naturally arises is thus: "What role do higher orders of co-occurrence play in LBD?" This question ties into the aforementioned theoretical work that we plan to undertake (Section 4). If statistics can be gathered for multiple orders of co-occurrence such as the number of connections, the weight of the connections, the importance of the connecting concepts, etc., then these statistics can be evaluated for use as attributes in instances in the training data. Naturally, other attributes would also need to be evaluated, but based on our experience, statistics characterizing higher-order co-occurrences may well play an important role.

There are concerns with this approach. The first is that the ratio of known, labeled instances (six of them!) to the search space (encompassing the entire Medline database) is vast. This highlights concerns common in machine learning, including overfitting the model, or not having sufficient data to train appropriately. Additionally, there are differences between even the six examples mentioned – differences which may prohibit standardization across them. The first two examples, Raynaud's↔Fish Oil and Migraine↔Magnesium, are similar because the connection path links a disease and a substance or dietary factor. One of their common points is that several of the efforts surveyed applied the open discovery process and replicated these connection discoveries. The Indomethacin↔Alzheimer's disease, Somatomedin C↔Arginine, and Schizophrenia↔Calcium-independent PhospholipaseA2 connections are examples of the closed discovery process because the first aim in these experiments was to find *B-concepts* linking between *A* and *C*. The last example, Therapeutic Uses for Thalidomide↔Four Diseases, differs from other examples in that the search space is restricted for a particular target. In this case, the authors selected only UMLS concepts classified as an "Immunologic factor" from sentences that co-occurred with Thalidomide because they hypothesized that they may find new therapeutic applications through the immunologic actions of the drug. Needless to say, obtaining additional training data is essential to the use of supervised machine learning algorithms for LBD.

It would seem that the way forward is hindered from the start because the development of training data requires novel discoveries to be made. No doubt such discoveries will emerge over time as LBD continues to grow in use, but in the near term it seems that the lack of suitable training data may prevent rapid progress in the application of supervised machine learning algorithms to LBD. We have to this point, however, limited the discussion to the medical domain simply because the available training instances are in this domain. The question we pose is "Would other domains provide a better opportunity for automation than the medical domain?" The following section addresses one such application in the patent domain. Because at present labeled instances in the medical domain are few, perhaps machine-learning-based automation may be more successful if training data can be more easily developed. Starting with a semi-automatic LBD system, it may be possible to develop a ground truth dataset sufficient to enable the application of a supervised machine learning algorithm. Naturally, the cross-domain scalability of a model so constructed would also need to be evaluated, but this may prove more tractable even for small training sets such as those currently available in the medical domain.

## 7. Hurdling the Barriers to Automating LBD using Supervised Machine Learning

One application domain that may be well served by the LBD process is the protection of intellectual property. Wu & Pottenger (2003) addresses information extraction in patents. A common scenario in this domain involves a researcher who has a technical problem and wishes to search a database to uncover patents addressing the problem, if any. This both minimizes the cost of "reinventing the wheel" as well as preventing possible resource-consuming litigation resulting from patent infringement. Since patent descriptions are generally complex and difficult to read, Wu & Pottenger (2003) propose a supervised learning algorithm that extracts Problem Solved Identifiers (PSIs) from the text. Each PSI is a sentence that describes an insufficiency in prior art. In developing training data in this approach, the training-set developer labels sentences of particular patents as PSIs or not. This training set is then used for the discovery of finite state automata, in the form of Reduced Regular Expressions, or RREs. Using the RREs so learned, one or more PSIs are extracted from each (previously unseen) patent during classification. Once PSIs have been extracted and stored, the resulting database of PSIs can be queried by researchers per the scenario described above.

What is the connection with LBD? Certainly LBD uncovers novel connections between disjoint datasets. In the patent data domain, there may be similar connections between distinct patent areas, much like the links between disjoint scientific literatures. When different patents are linked in this way they may yield valuable novel information. An example is linking a PSI to a solution to the problem other than the one detailed in the patent containing the PSI. In this scenario, the user begins with a technical problem and searches for a solution. The first step is thus to query the PSI database. Given that some matches occur, a number of patents are returned that address problems similar to the beginning problem. The user then studies patents in this result set to see if there exists an appropriate solution to the problem. If such a solution is found, then the process is complete. If not, however, then an LBD-like process can be applied to discover novel links between the user's original problem and potential solutions. In other words, the initial results returned by the PSI search system become the start literature, and the user proceeds to explore connections through intermediate concepts to target concepts representing solutions to the problem. These steps are summarized as follows:

1. Begin with a problem to be solved
2. Formulate the problem as a query to the PSI database
3. Study results returned to identify relevant documents that imply solutions
4. If no relevant documents are found in the results, begin the LBD process using the results as the start literature

In this way LBD may uncover novel solutions to problems that no single existing patent addresses. This is, by definition, knowledge discovery.

As an example, consider a researcher working on inkjet printers with a problem related to ink flow in the print head. Using traditional search mechanisms, the researcher searches for patents about ink flow in print heads within the computer hardware domain. Suppose no useful results are returned in this domain; still, a solution may exist in a different domain. One such domain might be fluid mechanics, where there may be patents that address a very different problem such as fluid flow in small pipes. This solution might be found to be relevant to the print head problem, as well.

The basis for considering the application of LBD in the patent domain is the supposition that developing a suitable ground truth data set will thereby be facilitated. The essential difference in the patent vs. the medical domains is that the aforementioned scenario involving patent search is widely used by numerous patent intelligence experts[12]. Unlike LBD in the medical domain, which is very limited in scope to research efforts, the (manual) identification of PSIs and

[12] In fact the authors of this survey are working closely with one such expert at a large US corporation.

subsequent searching for solutions is common in patent intelligence applications. In addition, the cost of verifying a novel link between a problem and a solution is negligible compared to that of verifying a link in the medical field. Thus the way forward becomes clear. Working with patent intelligence experts, a semi-automatic LBD system can be deployed that assists these experts in performing their information seeking tasks per the four steps outlined above. As part of this process, novel problem↔solution links will be discovered and catalogued, forming instances that become part of the ground truth needed to effectively apply supervised machine learning algorithms to LBD.

There is another practical reason to perform this research in the patent domain. Similar to the work surveyed in LBD that uses UMLS metadata for filtering, structures in the patent intelligence domain can be leveraged. Recall that one usage of UMLS was to eliminate concepts unrelated to the information sought. In the patent domain, there are several classification codes for patents such as the Derwent, United States, and international classification codes. Like thesauri in UMLS, these metadata structures are hierarchies that can be exploited by the LBD process to guide discovery in the patent domain.

## 8. Conclusion

Literature Based Discovery (LBD) has increasingly become the focus of research in knowledge discovery in textual data. The time is thus ripe for a survey that highlights the state of the art in this field. This work describes several LBD systems and provides detailed information on each regarding the input data and domain, how concepts are extracted from text, the approach and model of discovery, and how each system is evaluated. Additional discussion focuses on trends in approaches to LBD, and the key areas that are most in need of further work. The review of the literature indicates that although some progress has been made toward automating LBD, there is still much work to be done. Current systems continue to require manual processes, requiring human interaction at several points. Additionally, this survey finds that much of the research lacks formal evaluation metrics and methodologies to determine effectiveness. Understanding the underlying theory of LBD and developing effective metrics for evaluation is crucial for further progress in the field.

In summary, the areas of need identified in this survey include the need to develop theoretical foundations for LBD, a critical need to address evaluation methodologies and a need to more fully automate the LBD process. These three research thrusts are closely related, and this survey also provided direction needed to begin addressing these needs.

In conclusion, it is our contention that fully automated LBD is achievable employing supervised machine learning algorithms, but the research must be conducted in a domain other than the field of medicine. It is our intention to explore these issues as part of our continuing research in information extraction and text mining in the Parallel and Distributed Text Mining Lab at Lehigh University.

**Acknowledgements**

This work was supported in part by National Science Foundation grants DGE-0231768 and IIS-0534276. The authors wish to acknowledge Dr. Daniel Phelps of Eastman Kodak Company and Mr. Tianhao Wu of Ask Jeeves, Inc. for their ongoing support of and collaboration with the research in information extraction.  Co-authors William M. Pottenger and Christopher D. Janneck also wish to thank their Lord and Savior, Jesus the Christ, for the salvation of their lives. In Jesus' Name, Amen.

Table 1: Summary of LBD Research

**References**

Agrawal, R., Mannila, H., Srikant, R., Toivonen H., & Verkamo, A.I. (1995). Fast Discovery of Association Rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth,  R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining. (pp. 307-328). AAAI/MIT Press.

Aronson, A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of AMIA Symposium. (pp. 17-21).

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 34–43

Cory, K.A. (1997). Discovering hidden analogies in an online humanities database. Computers and the Humanities, 31(1), 1-12.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391-407.

Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. New York: Wiley-Interscience Press.

Frank, E., Paynter, G. W. , Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-Specific Keyphrase Extraction. Proceedings of Sixteenth International Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers.

Gordon, M.D., & Dumais, S. (1998). Using Latent Semantic Indexing for literature based discovery. Journal of the American Society for Information Science and Technology, 49(8), 674-685.

Gordon, M.D., & Lindsay, R.K. (1996). Towards discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science, 47(2), 116-128.

Gordon, M., Lindsay, R.K., & Fan, W. (2001). Literature Based Discovery on the World Wide Web. ACM Transactions on Internet Technology, 2(4), 261-275.

Hamosh, A., Scott, A.F., Amberger, J., Valle, D., & McKusick, V.A. (2000). Online Mendelian Inheritance in Man (OMIM). Human Mutatation, 15, 57-61.

Hristovski, D., Stare, J., Peterlin, B., & Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in MEDLINE and UMLS. Medinformation, 10(2), 1344-1348.

Hristovski, D., Peterlin, B., Mitchell, J.A., & Humphrey, S.M. (2005). Using literature-based discovery to identify disease candidate genes. International Journal of Medical Informatics, 4(2-4), 289-98.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding LSI performance. Information Processing & Management, 42(1), 56-73.

Lindsay, R.K., & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50, 574-587.

Lindberg, D.A.B., Humprey, B.L., & McCray, A.T. (1993). The unified medical language system. Methods of Information in Medicine, 32, 281-291.

Lowe, H.J., & Barnett, G.O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. Journal of the American Medical Association, 271, 1103-1108.

Maglott, D.R., Katz, K.S., Sicotte, H., & Pruitt, K.D. (2000). NCBI's LocusLink and RefSeq. Nucleic Acids Research, 28, 126-128.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., & Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). Human Genetics, 109, 678-680.

Pratt, W., & Yetisgen-Yildiz, M. (2003). LitLinker: Capturing Connections across the Biomedical Literature, Proceedings of the International Conference on Knowledge Capture (K-Cap'03). Florida, October 2003.

Rikken, F., & Vos, R. (1995). How adverse drug reactions can play a role in innovative drug research. Pharmacy World and Science, 17(6), 195–200.

Skeels, M.M., Henning, K., Yetisgen-Yildiz, M., & Pratt, W. (2005). Interaction Design for Literature-Based Discovery. Proceedings of the ACM International Conference on Human Factors in Computing Systems (CHI 2005). Portland, OR. April 2005

Smalheiser, N.R., & Swanson, D.R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. Neuroscience Research Communications: 15(1), 1-9.

Smalheiser, N.R., & Swanson, D.R. (1996a). Indomethacin and Alzheimer's disease. Neurology, 46, 583.

Smalheiser, N.R., & Swanson, D.R. (1996b). Linking estrogen to Alzheimer's Disease: An informatics approach. Neurology, 47(3), 809-810.

Smalheiser, N.R., & Swanson, D.R. (1998a). Calcium-independent phospholipase A2 and Schizophrenia. Archives of General Psychiatry. 55(8), 752-753.

Smalheiser, N.R., & Swanson, D.R. (1998b). Using ARROWSMITH: A computer assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine, 57, 149–153.

Srinivasan, P. (2004). Text Mining Generating Hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology, 55(5),  396-413.

Srinivasan, P., & Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics, 20(1), 290-296.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7-18.

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Medicine, 31(4), 526-557.

Swanson, D.R. (1990). Somatomedin C and arginine: Implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine, 33(2), 157-186.

Swanson, D.R. (1991). Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.), Proceedings of the 14th Annual

International ACM/SIGIR Conference on Research and Development in Information Retrieval (pp. 280–289). New York: ACM Press.

Swanson, D.R. (2001). On the fragmentation of knowledge, the connection explosion, and assembling other people's ideas. ASIST Award of Merit acceptance speech, Bulletin of ASIST 27(3), 12-14.

Swanson D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures:  a stimulus to scientific discovery. Artificial Intelligence, 91, 183-203.

Swanson, D.R., Smalheiser, N.R., & Bookstein, A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. Journal of the American Society for Information Science, 52(10),  797-812.

Valdes-Perez, R.E. (1994). Conjecturing hidden entities by means of simplicity and conservation laws: Machine discovery in chemistry. Artificial Intelligence, 65(2), 247-280.

Van der Eijk, C., Van Mulligen, E., Kors, J.A., Mons, B., & Van den Berg, J. (2004). Constructing an associative concept space for literature-based discovery. Journal of the American Society for Information Science and Technology, 55(5), 436-444.

Weeber, M. (2004). Advances in Literature-Based Discovery. http://nl.ijs.si/et/talks/tsujiilab/saso/weeber.pdf

Accessed August 29, 2005.

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong-van den Berg, L.T.W., & Vos, R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. In J.M. Overhage (Ed.), Proceedings of the 2000 AMIA Annual Fall Symposium (pp. 903–907). Philadelphia, PA: Hanley and Belfus.

Weeber, M., & Molema, G. (2004). Literature-based discovery in biomedicine. http://math.nist.gov/~JDevaney/CommKnow/mar2001/weeber.stanford.ppt
Accessed August 29, 2005.

Weeber, M., Vos, R., Klein, H., & de Jong-van den Berg, L.T.W. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud – fish oil and Migraine – magnesium discoveries. Journal of the American Society for Information Science and Technology, 52(7), 548-557.

Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L.T.W, Aronson, A., & Molema, G. (2003). Generating hypothesis by discovering implicit associations in the literature:A case report for potential therapeutic uses for Thalidomide. Journal of American Medical Informatics Association, 10(3), 252-259.

Witten, I.H., & Frank, E. (2000). Data Mining: Practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann.

Wu, T. & Pottenger, W. M. (2003). A Supervised Learning Algorithm for Information Extraction from Textual Data. In M.W. Berry and W.M. Pottenger (Eds.), Proceedings of the Workshop on Text Mining, Third SIAM International Conference on Data Mining, (pp. 60-71), San Francisco, CA. SIAM Press, Philadelphia, PA.

Wu, T., Holzman, L. E., Pottenger, W. M., & Phelps, D. J. (2003). A Supervised Learning Algorithm for Information Extraction from Textual Data. Proceedings of the Textmine '03 Workshop, Third SIAM International Conference on Data Mining, San Francisco, CA, May

Wren, J.D., & Garner, H.R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods of Information in Medicine, 41, 426-434.

Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., & Garner, H.R. (2004). Knowledge Discovery by Automated Identification and Ranking of Implicit Relationships. Bioinformatics, 20(3), 389-98.

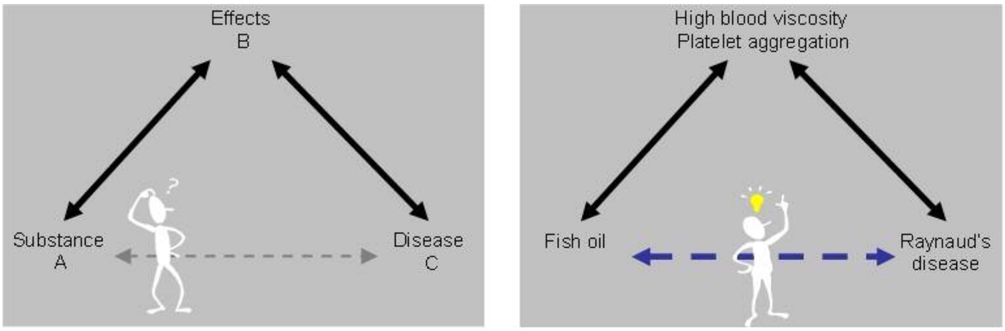


**Figure 1: Swanson's discovery: Raynaud's↔Fish Oil connection from (Weeber, 2001)** [2]

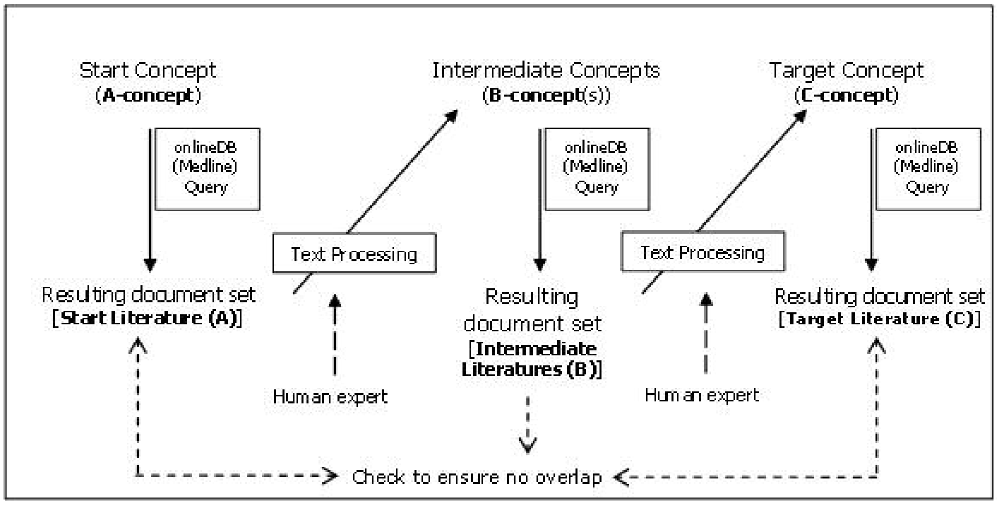


**Figure 2: Swanson's framework, which is followed by many other works (Open Discovery)**

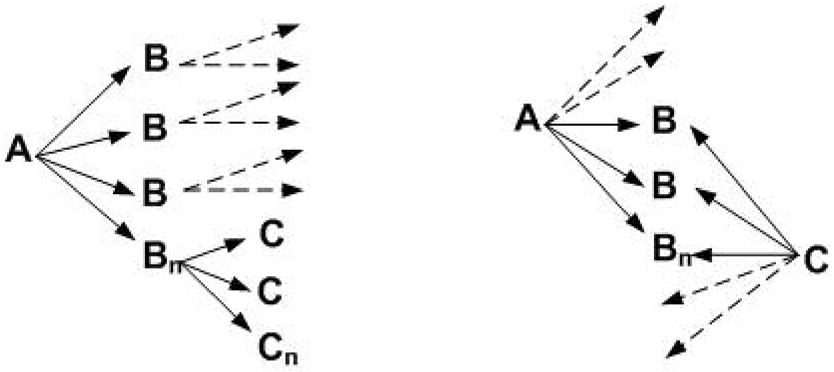


**Figure 3: Open Discovery vs. Closed Discovery. Adapted from (Weeber et al., 2001). □The solid arrows indicate potentially interesting pathways of discovery, the dashed ones unsuccessful pathways.□**

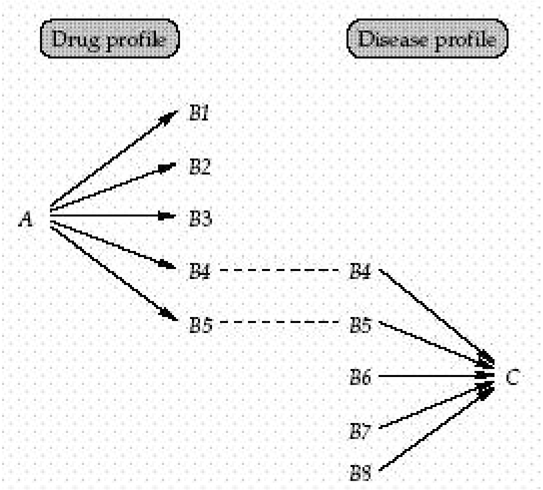


**Figure 4: □Vos's and Swanson's model of discovery combined.□ (Weeber, 2004)**

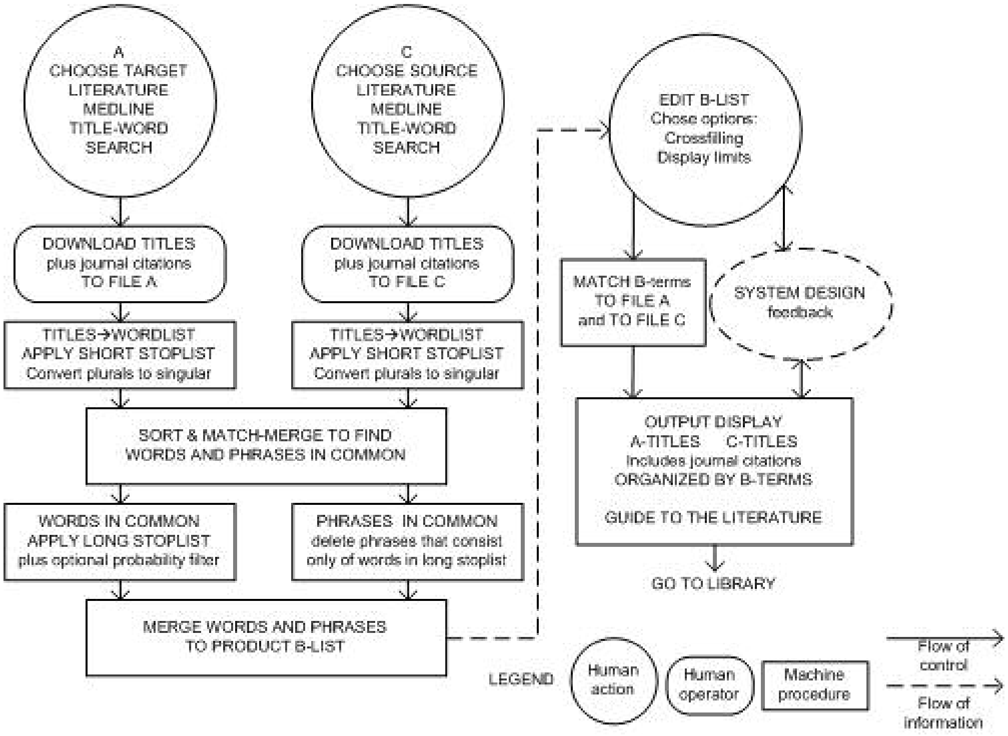


**Figure 5: A flowchart of closed discovery process. Adapted from Swanson & Smalheiser (1997)**

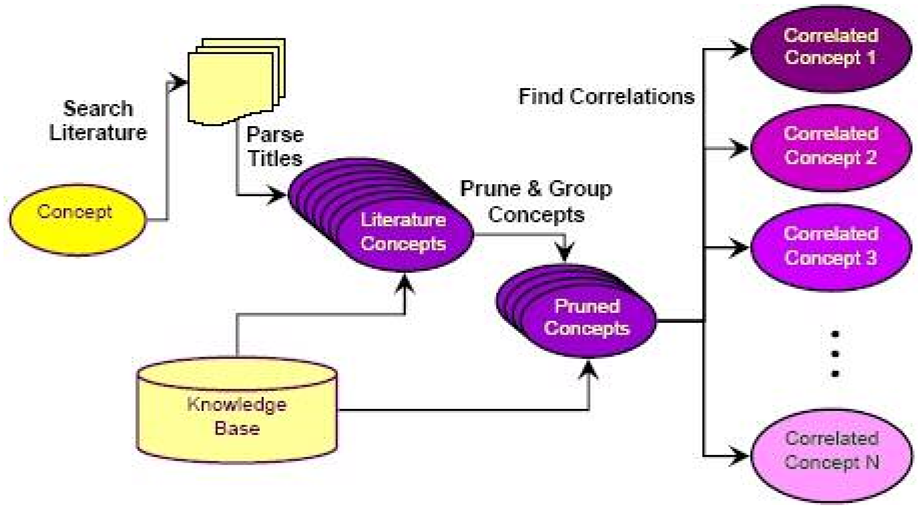


**Figure 6: The Text Mining Process in LitLinker (Pratt & Yetisgen-Yildiz, 2003)**

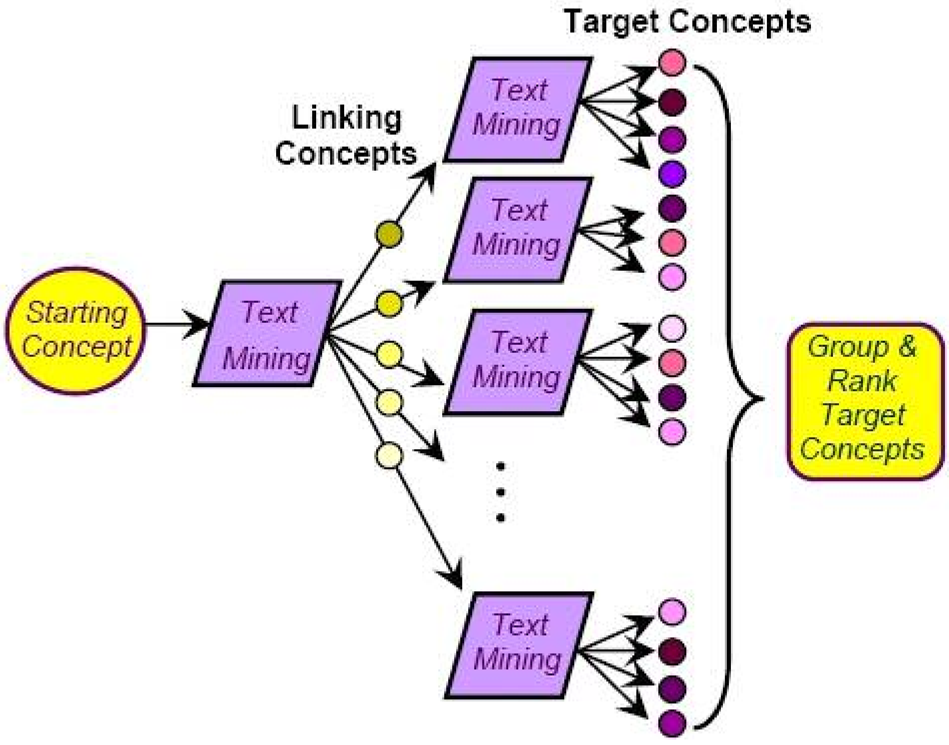


**Figure 7: Discovery Process in LitLinker (Pratt & Yetisgen-Yildiz, 2003)**

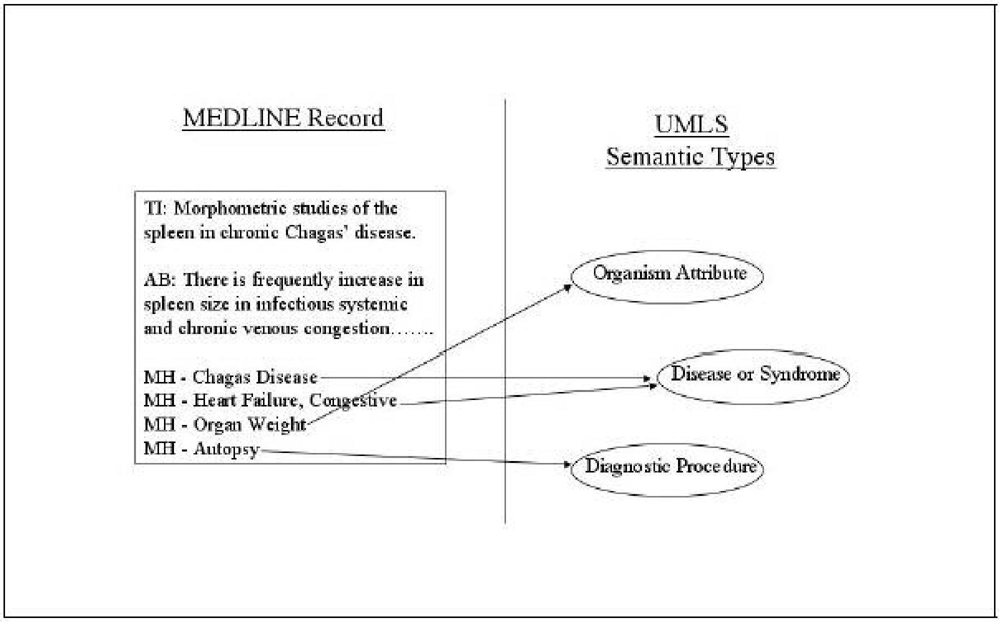

MEDLINE Record
UMLS
Semantic Types
TI: Morphometric studies of the spleen in chronic Chagas' disease.
AB: There is frequently increase in spleen size in infectious systemic and chronic venous congestion…….
MH - Chagas Disease
MH - Heart Failure, Congestive
MH - Organ Weight
MH - Autopsy
Organism Attribute
Disease or Syndrome
Diagnostic Procedure

**Figure 8: Semantic types of MeSH terms (Srinivasan, 2004).**

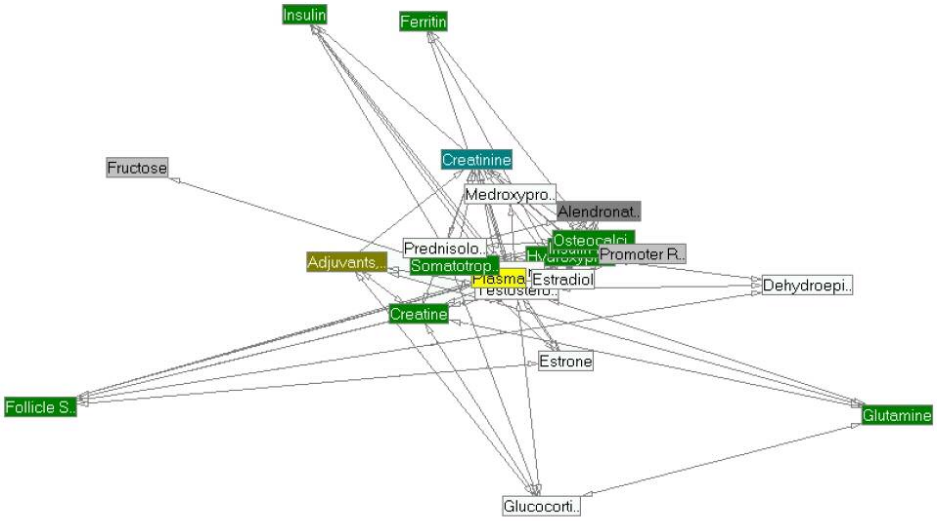


**Figure 9: Two-dimensional projection of part of an 8-dimensional ACS from a set of Medline abstracts on Muscular Dystropy. The ACS suggests a relationship between □Insulin□ and □Ferritin□ (Van der Eijk et al., 2004).**

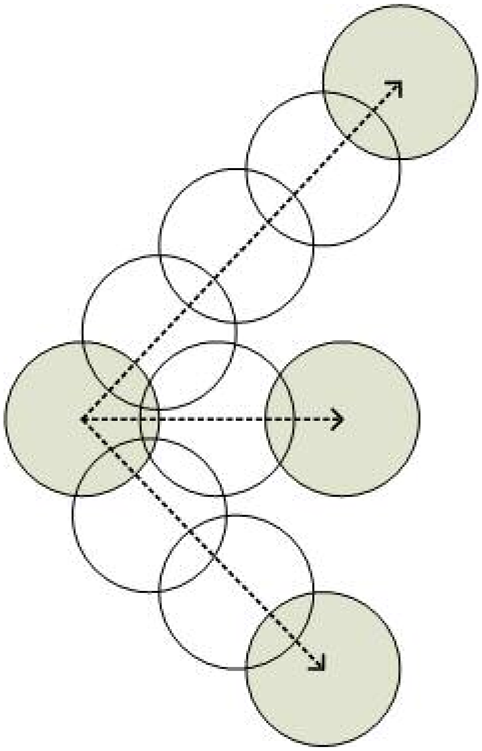
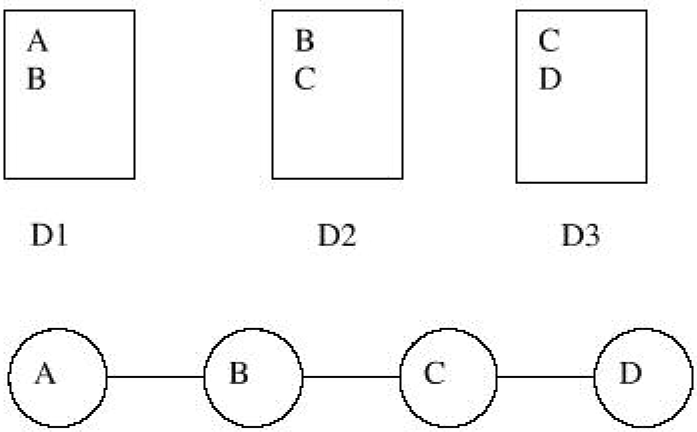


**Figure 10: Multiple search paths with multiple intermediate literatures. Adapted from (Gordon & Lindsay, 1996). Examining many B-concepts may reveal many others and carry us to different targets.**



1st order term co-occurrence {A,B}, {B,C}, {C,D}
2nd order term co-occurrence {A,C}, {B,D}
3rd order term co-occurrence {A,D}

**Figure 11: Higher order co-occurrences (from Kontostathis & Pottenger, 2006)**

Table 1: Summary of LBD Research

| Authors | Years | Domain & Database | Input Data | Model(s) of Discovery | Approach | Discoveries | Evaluation |
|---|---|---|---|---|---|---|---|
| Swanson | 1986<br>1988<br>1991 | Medical – Medline | Titles of Medline records | - Open discovery (first, mostly manually)<br>- Closed discovery (in later work) | - Co-occurrence analysis for title words and phrases<br>- A→B→C | - Fish Oil & Raynaud's syndrome<br>- Migraines & magnesium | These novel connections are later proved by publications of medical researchers |
| Swanson & Smalheiser | 1994<br>1996<br>1997 | Medical – Medline | *Original:*<br>- Titles of Medline records<br>- Domain-specific stop-word lists<br>*Added in Web version:*<br>- Medline's MeSH headings<br>- UMLS Semantic categories (filtering) | - Open discovery (in the 1997 paper)<br>- Closed discovery (in Web version) | - Low document frequency (df) terms eliminated<br>- Number of links used to identify targets<br>- A→B→C | - Magnesium deficiency & neurological disease<br>- Indomethacin & Alzheimer's Disease<br>- Estrogen & Alzheimer's Disease | Compare to the original Migraine↔Magnesium and Raynaud's↔Fish Oil connections |
| Gordon & Lindsay | 1996 | Medical – Medline | - Complete Medline records<br>- Every term or adjacent bigram phrase<br>- Three different stop-word lists | Open discovery | *Statistics:*<br>- Token frequency<br>- Record frequency<br>- Token frequency*log-inverse global record frequency<br>*Then:*<br>- Identify linking concepts based on high occurrence frequency<br>- A→B→C | Fish Oil & Raynaud's syndrome | - Compare to Raynaud's–Fish Oil connection<br>- Use precision and recall on linking concepts |
| Gordon & Dumais | 1998 | Medical – Medline | - Complete Medline records<br>- Every term or adjacent bigram phrase<br>- Three different stop-word lists<br>(*similar to previous work*) | Open discovery | - Latent Semantic Indexing (LSI)<br>- A→B→C | Fish Oil & Raynaud's syndrome | Intersection found with top concept lists from Gordon & Lindsay (1996) |
| Gordon & Lindsay | 1999 | Medical – Medline | - Complete Medline records<br>- Every term or adjacent bigram and trigram phrase<br>- Three different stop-word lists | Open discovery | *Lexical analysis using:*<br>- Token count<br>- Document counts<br>- Relative frequency<br>- tf*idf<br>- A→B→C | Migraines & magnesium | Compare to the Migraine↔Magnesium connection |
| Weeber, Vos, Klein & de Jong-van den Berg | 2001 | Medical – Medline | - UMLS concepts acquired from titles and abstracts<br>- UMLS Semantic categories (filtering) | - Open discovery<br>- Closed discovery | - Filter for concepts that co-occur with either A or C in a sentence<br>- Use concept frequency in open approach<br>- Use number of links in closed approach<br>- A→B→C | - Fish Oil & Raynaud's syndrome<br>- Migraines & magnesium | Compare to the original Migraine↔Magnesium and Raynaud's↔Fish Oil connections |

| Authors | Years | Domain & Database | Input Data | Model of Discovery | Approach | Discoveries | Evaluation |
|---|---|---|---|---|---|---|---|
| Gordon, Lindsay & Fan | 2001 | WWW via AltaVista search engine | Top 100 web pages returned by a query composed of that concept | Open discovery | *Statistics:*<br>- Token frequency<br>- Document frequency<br>- A→B→C | Potential novel application areas for “genetic algorithms” | No formal evaluation |
| Pratt & Yetisgen-Yildiz | 2003 | Medical – Medline | - UMLS concepts acquired from only title text<br>- UMLS Semantic categories (filtering) | Open discovery | - Prune concepts that are too general or too close to start term using UMLS hierarchy<br>- Cluster related concepts<br>- Use ARM algorithm to identify concept correlations<br>- A→B→C | Migraines & magnesium | Compare to the Migraine↔Magnesium connection |
| Srinivasan | 2004 | Medical – Medline | - Medical Subject Heading (MeSH) terms in Medline records<br>- UMLS Semantic categories (filtering) | - Open discovery<br>- Closed discovery | - Build topics using MeSH-based profiles from Medline<br>- Use tf*idf for weighting<br>- A→B→C | Ranks key MeSH terms in top-10 for 2 open and 5 closed prior discoveries | Compare top ranked terms to many connections, including: Raynaud↔Fish Oil, Migraine↔Magnesium |
| Van der Eijk, Van Mulligen, Kors, Mons & Van den Berg | 2004 | Medical – Medline | - Abstracts of Medline records (with stop-word elimination)<br>- MeSH thesaurus (mapping medical concepts) | Visualization for open or closed discovery: concepts that are connected by frequent co-occurrence paths, either directly or indirectly have a small distance in resulting space | - Build fingerprints from term frequencies and depth in the thesaurus’ hierarchy<br>- Form ACS using co-occurrence of concepts in fingerprints | None | Use simulated scientific literature data sets for which the outcome is predictable |
| Wren, Bekeredjian, Stewart, Shohet & Garner | 2004 | Medical – Medline | - Titles and Abstracts of all electronically-available Medline records<br>- Multiple reference databases (OMIM, MeSH, LocusLink, HGNC)<br>- Acronym Resolving General Heuristic | Open discovery | - Find implicit A-C relationships by co-occurrence<br>- Compare relationships against a random network model<br>- Return top concepts with most connections<br>- A→B→C | Cardiac Hypertrophy & Chlorpromazine | Perform clinical study (using mice) to test generated hypothesis |
| Hristovski, Peterlin, Mitchell & Humphrey | 2005 | Medical – Medline | - Titles and Abstracts of Medline records<br>- MeSH descriptors<br>- LocusLink and HUGO (gene mapping) | Open discovery | - Find A-C relationships through association rules<br>- A→B→C | Potential method for gene symbol disambiguation | No formal evaluation |