1-1-2006

Literature based discovery: Techniques and tools

Ramalakshmi Sundar
University of Nevada, Las Vegas

LITERATURE BASED DISCOVERY: TECHNIQUES AND TOOLS

by

Ramalakshmi Sundar

Bachelor of Engineering in Electronics & Communication
University of Madras, India
May 2003

A thesis submitted in partial fulfillment of the
requirements for the

Master of Science Degree in Computer Science
School of Computer Science
Howard R. Hughes College of Engineering

Graduate College
University of Nevada, Las Vegas
May 2007

UNLV

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

MARCH 15TH , 2007

The Thesis prepared by

RAMALAKSHMI SUNDAR

Entitled

LITERATURE BASED DISCOVERY: TECHNIQUES AND TOOLS

is approved in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT

Literature Based Discovery: Techniques and Tools

by

Ramalakshmi Sundar

Dr. Kazem Taghva, Examination Committee Chair
Professor of Computer Science
University of Nevada, Las Vegas

Literature Based Discovery (LBD) was initially proposed by Don R. Swanson in 1980 as a method to establish relationships between disease and remedy from disjoint science literature. Consequently, he established a link between magnesium and migraines. Since then literature based discovery has been a subject of research and development for discovery in online medical publications. It has further been investigated in both chemistry and mathematics.

In this thesis, we give an overview of LBD and the software tools necessary to automate this technique. We further provide an implementation of this technique that is intended to be used for computer science subject matter.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. My first, and most earnest, acknowledgment must go to my Advisor and Chair of my Committee Dr. Kazem Taghva. I have been indebted in the preparation of this thesis to my Advisor, whose patience and kindness, as well as his academic experience, has been invaluable to me.

I am extremely grateful to Dr. Thomas Nartker, Dr. Shahram Latifi and Dr. Ajoy Datta for being a part of my Thesis Advisory Committee. I also owe a huge debt of gratitude to Dr. Ajoy Datta for his continuous guidance, advice and encouragement throughout my academic years. I am very grateful to my friend Karthik Raghavan, for extending his valuable time to solve critical issues in my programming tool.

I have been extremely fortunate to have the support of very special friends to whom I am truly grateful. My final and most heartfelt acknowledgment goes to my parents and my brother, on whose constant encouragement and love I have relied throughout my time at the University.

CHAPTER 1

INTRODUCTION

1.1. Literature Based Discovery

In a world with seemingly boundless increase 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. Dealing with a substantial amount of information has led to fragmentation of scientific literature. Scientists tend to correspond more to fragments than with the field's broader community. Results are published, yet 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 computer-aided literature searching or Information Retrieval, are insufficient for recognizing connections between fragments. A solution to the problem is Literature-Based Discovery (LBD) which directly addresses the problems of knowledge overspecialization [3]. Literature-based discovery of knowledge is an important addition to the knowledge worker's repertoire of skills, and is separate and distinct from data mining and information retrieval [1].

The goal of literature-based discovery in general is to discover new, potentially meaningful relations between a given starting concept of interest and other concepts, by mining bibliographic databases [25]. Recently, several LBD tools have been developed

and a few well-motivated, specific and directly testable hypothesis have been published some of which have been validated experimentally [8].

The literature-based discovery process can be diagrammed like this:

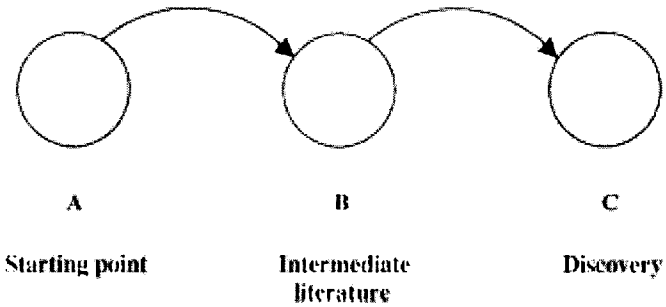

Figure 1.1: Literature-Based Discovery Model

What has been constant, though, are several features in undertaking the task of literature based discovery [10]:

- a *scientific* literature as the source of discovery;
- discoveries have generally proceeded from *disease* to *cure* (problem to solution);
- the analysis of two or more literatures has been necessary to produce a discovery; and
- what constitutes a literature-based discovery has been a connection either *completely new* to the field or at least *overlooked by the vast majority* of its practitioners.

If two literatures are linked by arguments that they respectively put forward-that is, are "logically" related or connected-one would expect them to cite each other. If they do not, then the logical connections between them would be of great interest, for such

connections may be unintended, unnoticed, and unknown-therefore potential sources of new knowledge [37].

1.2. Undiscovered Public Knowledge

Undiscovered public knowledge addresses bodies of information that are similar but distinct or not normally connected. *Public* because all pieces of knowledge needed already exists and are publicly available, *undiscovered* because no scientist has brought the pieces together yet. Using bibliographic analysis, it is possible to find links between published information that had not previously been known by researchers. This idea was first advanced by Don Swanson of the Chicago Library School [4].

A more complex, and perhaps more interesting, example will further illuminate the idea of undiscovered public knowledge. Suppose the following two reports are published separately and independently, the authors of each report being unaware of the other report: (i) a report that process A causes the result B, and (ii) a separate report that B causes the result C. It follows of course that A leads to, causes, or implies C. If the two reports, i and ii have never together become known to anyone, then we must regard "A causes C" as an objectively existing but as yet undiscovered piece of knowledge – a missing link [2].

Swanson labeled as "complementary but disjoint" topical literatures that are potentially combinable toward some new end, but that are not already bibliographically related. For an example Swanson discovered that two bodies of literature – one on the circulatory effects of dietary fish oil and other on the circulatory disorder of Raynaud's disease – had no direct connection (i.e., no researcher has yet used fish oil to treat

Raynaud's disease), but did suggest a connection worth exploring through this unique bibliographic analysis [4].

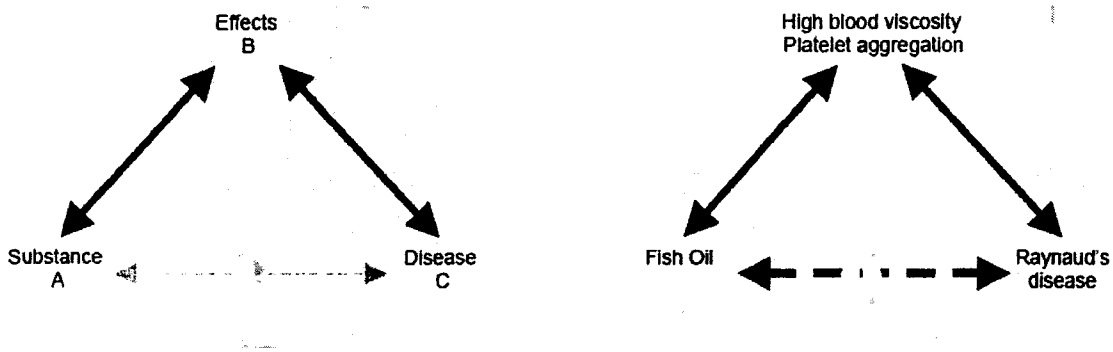

Figure 1.2: Swanson's Undiscovered Public Knowledge

Swanson suggests that there are many other disconnected fragments of knowledge in the literature to which his analysis would be able to make connections. Swanson has shown how chains of causal implication within the medical literature can lead to hypotheses for causes of rare diseases, some of which have received supporting experimental evidence [7]. Utilizing scientific literature as a potential source of new knowledge is an extremely attractive idea. Hence undiscovered public knowledge is related to the larger problem of "literature based discovery", wherein the body of research literature is employed as a source of knowledge by researchers [4, 5].

1.3. Goal, Methodology and Structure of this thesis

Goal

The main goal of this thesis is to make recommendations on a system that helps scientists and investigators to make discoveries in undiscovered public knowledge in

scientific literature. We focus on developing an interactive search engine based tool that helps the user to find novel, implicit, hidden connections between literatures

Methodology

We will begin our work by studying some of the available tools and techniques. Since the field of biomedicine has been of great interest in literature based discovery, we shall pay attention to this domain. We will then examine and compare the existing systems to aid LBD in life sciences and from this comparison; we will derive guidelines for our own system.

We will give an overview of two basic approaches to searching for undiscovered public knowledge. We will also suggest possible improvements to these approaches.
Our approach in the LBD will be based on a preliminary Search tool that is implemented in RUBY. We will show how this tool can be used to mimic Swanson's technique

Structure

The structure of this thesis is as follows: Chapter 2 is about the literature review of the LBD process, which extensively describes Swanson's methodology. In Chapter 3 we give a detailed description of the available LBD techniques and tools and a description of the popular Arrowsmith system developed by Don Swanson in an attempt to automate the LBD process. Chapter 4 focuses on the implementation of our LBD tool, which is basically an effort to replicate Swanson's Arrowsmith system. Chapter 5 is the experimental evaluation of our tool and enhancements. Chapter 6 will be our conclusion and future work.

CHAPTER 2

LITERATURE REVIEW OF LBD

Literature-based discovery is a discipline within the information sciences domain that tries to discover new knowledge by combining existing knowledge which is written down in scientific literature [17]. Employing techniques from Information Retrieval and Natural Language Processing, LBD has potential for widespread application yet is currently implemented primarily in the medical domain [3]. LBD mainly comprised of two stages, getting from the problem to a conjectural solution and exploring in depth the connections between the two disjoint literatures. First, literature-based discovery uses as its input collections of ordinary documents and not transaction logs, database relations, or other structured information. Second (and most important), LBD seeks relationships that, by definition, are not contained within an existing textual corpus, unlike efforts that seek correlations, patterns, or rules within defined set of texts [10].

2.1. Swanson Linking (SL)

Swanson Linking (SL) is a method which tries to disclose "hidden" (i.e. unpublished) but implicit links between concepts not mentioning each other in their respective literature representations. The SL analysis may give rise to the development of new hypotheses. The published examples of this kind of "literature-based discovery" involve basically three different sets of literature [31]:

(i) A problem-based literature – e.g. describing a disease – is referred to as "source"; (ii) A literature not being mentioned in the source literature but possibly contributing to problem solving is called "target"; (iii) A literature representing a major concept which is relevant for and occurs in both, source and target literature, is labeled "intermediate "(Swanson and Smalheiser (1999)) [31].

The discovery process might normally proceed from source to target via intermediate; however, the reverse order is naturally conceivable, and any coherent literature set regarded as "intermediate" may be explored for source and target concepts simultaneously [31].

2.2. Swanson's ABC Model

In 1986, Don Swanson presented his first literature-based hypothesis that fish oil may have beneficial effects in patients with Raynaud's disease. Fish oil lowers blood viscosity, inhibits platelet aggregation and causes vascular reactivity. On the other hand, patients with Raynaud's disease have increased blood viscosity and platelet aggregation and suffer from impaired vascular reactivity [26].

In 1986, no one had made this implicit link explicit until Swanson connected the apparently disconnected fields of biomedical expertise. Swanson, has indeed, opened up a new way of doing information science, a new avenue for approaching information [26]. The possibility of linking different scientific disciplines through intermediate, or shared, interests has commonly been described as Swanson's *ABC* model (Figure 2.1) [8].

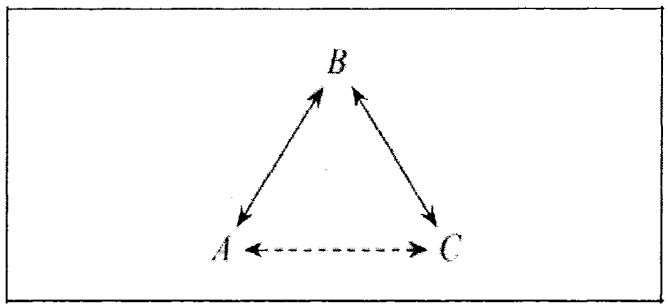

Figure 2.1: Swanson's ABC model of discovery

The *ABC* model of complementarity describes that, if one area of literature shows that *A* is connected to *B* and a different area shows that *B* is related to *C*, then bringing together these two areas for the first time may suggest a novel hypothesis that connects *A* and *C*, an implicit but not explicit connection[13]. Complementary refers to the relationship between two separate scientific arguments which, when combined, yield important inferences and insights not apparent in the separate arguments [42]. Complementartity does not necessarily imply logical transitivity, but rather is used in the looser sense of suggestibility [13].

The possibility of LBD implied by the above model underscores two important properties of sets of scientific articles- complementarity and noninteractivity. Two sets of articles are defined here as complementarity if together they can reveal useful information not apparent in the two sets considered separately [13].; "Noninteractive" means that the two pairs have no articles in common, do not cite each other, and are not co-cited) thus implying any logical relationships between them may be unintended and perhaps unnoticed [41]

2.3. A Two-Step Model of Discovery

Swanson's first discovery was a coincidence. By reading two different literature (for two different purposes), he made a connection between these literatures and formulated the hypothesis that fish oil may be used for treating Raynaud's disease. The potential of Swanson's research has been widely acknowledged, but likewise its complexity (Hearst, 1999). This complexity concerns the vast information space and possible number of connections [11].

Swanson and his co-worker Smalheiser use MEDLINE (NLM, 2000a) as their bibliographic database, with over 10 million citations of publications covering almost every scientific discipline in biomedicine. An addition to this complexity is that most information has been represented by natural language; therefore Natural Language Processing (NLP) techniques are needed to tackle the variation and intricacies of natural language. Swanson was able to validate this hypothesis by exploring MEDLINE extensively and by reading many scientific articles. Figure 2.2 shows a Venn diagram of his argument [11].

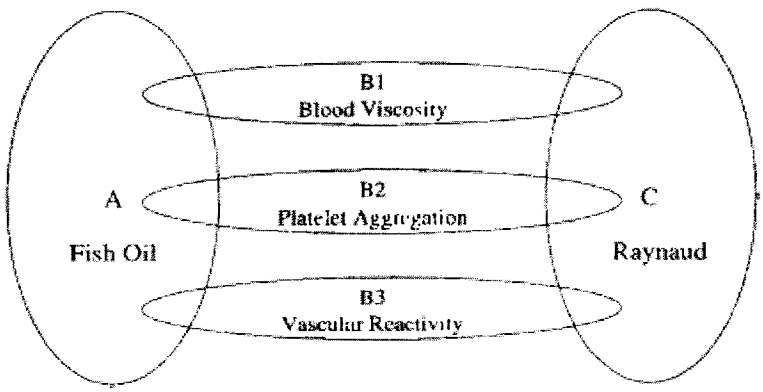

Figure 2.2: Venn diagram representing Swanson's first discovery

Already, in Swanson's first discovery, we may observe a two-step approach in the actual discovery process. A hypothesis has to be formulated or generated that may subsequently be validated or tested by extensive bibliographical analysis. The hypothesis can be generated in many not necessarily text-based ways [11].

Testing a hypothesis means assessing its plausibility. In his early discoveries, Swanson had formed, in one way or another, a hypothesis that he subsequently tested by extensive literature search and analysis. In later research, Swanson developed a method to also generate the hypothesis by bibliographic analysis (Swanson, 1991) [11].

Swanson's ABC model can be implemented in two different discovery processes.

a. Open Discovery Process - Characterized by the generation of a hypothesis

b. Closed Discovery Process – Characterized by the elaboration of a hypothesis

2.3.1. Open Discovery Process

Figure 2.3 depicts the open approach starting with disease C. Interesting clues (B) about the mechanism of the disease will be sought in the literature. In terms of Swanson's first discovery, the problem is to find underlying physiological mechanisms of Raynaud's disease. For the most interesting clues, substances (A) are looked for that may interact with these mechanisms. Swanson focused on dietary factors that may have an influence on the relevant B Processes. In the discovery process, it is likely that many Bs and As will be found. As the result of an open discovery process, one may formulate the specific hypothesis that substance A can be used for the treatment of disease C via one or more B pathways [8].

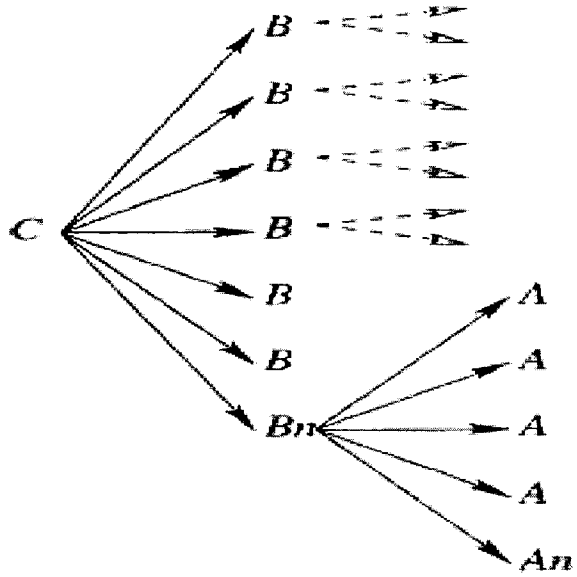

Figure 2.3: Open Discovery Process

2.3.2. Closed Discovery Process

Figure 2.4 depicts the closed discovery approach of verifying and elaborating an initial hypothesis, for instance, the treatment of disease C with substance A. Information on common mechanistic processes (B) are extracted from the literature. Typically, the more pathways between A and C are extracted, the stronger the hypothesis will be. An example relates to the observation that in patients with multiple myeloma who were treated with thalidomide, two responders had a concomitant improvement in chronic hepatitis C. A closed discovery process may elucidate possible underlying mechanisms of how thalidomide may treat chronic hepatitis C [8].

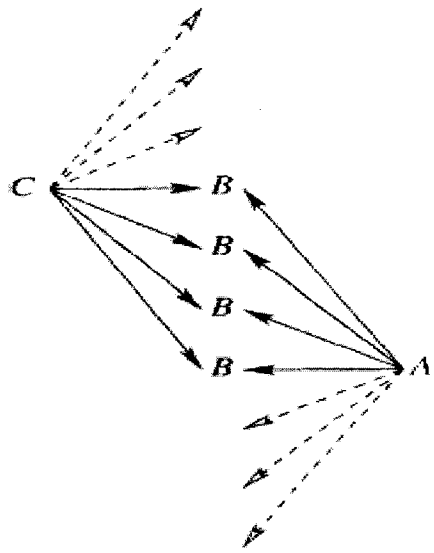

Figure 2.4: Closed Discovery Process

2.4. Direct Versus Indirect Connections between two Literatures

To find scientific literature on the relationship between any two searchable terms such as a chemical substance (A) and a disease (C), one could search Medline (Medical database) for all the records on A, all records on C, then form the intersection of Set A with Set C. This straightforward procedure is called here a search for "direct" A-C connections. (A or C can refer either to a search term or to a set of records created by the search). If there are no articles within the AC intersection, an intriguing question arises as to whether there might exist implicit or indirect connections between A and C based, for example, on an intermediate literature (B) for which both an AB relationship and a BC relationship have already been separately reported, but perhaps not considered together [9].

It is usually pointless or at least inefficient to pursue indirect connections before conducting a broadly based literature search for direct connections and analyzing the

relevant findings. Any direct connection that exits may eliminate or greatly change the problem and significance of indirect connections [9].

2.5. Data Mining in Datasets of keywords that signify Concepts

[1] This is a method for inductive generation of new concepts from analyzing existing concepts. Swanson has extensively tested one particular approach for finding undiscovered public knowledge. The plan is to discover implicit links between the topics A and C.

In brief, the methodology is statistical analysis of title keywords from 'complementary literatures', assuming

- Existence of an intermediate concept B (which may be conjecture initially) between a concept represented in a set A of articles and a concept represented in a second set C of articles, where
- this concept can be deduced from putative co-citation of keywords B in sets A and C, and
- hypotheses may be generated and/or substantiated through examination of documents retrieved using keyword sets AB and BC.

Since the method operates on title, the analysis draws on either the lexical content of the full text item or its title metadata. In terms of the six types of undiscovered public knowledge listed earlier, it belongs to the second category, of inference from transitive relations, since the keywords are assumed to act as markers of partial syllogisms AB and BC, thus inferring AC [10].

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, 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 [3].

2.6. Hypothesis Generation

Hypothesis generation, a crucial initial step for making scientific discoveries, relies in prior knowledge, experience, and intuition [36]. Practically all of the work in hypothesis generation makes use of an idea originated by Swanson in the 1980s called the 'complementary structures in disjoint literatures' (CSD). While Swanson applied his *ABC* model manually, several investigators have tried to automate the process [23].

Automated hypothesis generation systems may generate potential hypotheses, and therefore some method of evaluating these systems is necessary. It remains to be seen whether these systems will address the needs of the biomedical research community, but it is clear that they are a step towards addressing the needs of researchers beyond those served by search engines such as PubMed and Google [23].

2.7. Aim of LBD Systems

Literature-based discoveries are concise, well motivated and testable: a specific *C* is connected to a specific *A*. In reaching this specificity, intermediate steps are often not that precise. For instance, when looking for genes (typical *A*s), involved in a specific disease *C*, the actual approach is a hybrid of a closed and open discovery. This aspect may be used to conduct an open-like search with tools that only support closed discoveries [8].

Literature-based discovery relies on information retrieval-based techniques and insights, but is a much harder problem. Whereas information retrieval has, at the outset, the objective of finding documents relevant to a given need for information, the success of literature-based retrieval depends on finding topics (or documents) that are only indirectly relevant to the topic one uses to initiate the discovery process. In addition, what is found must be previously unknown in relation to the starting point [10].

Literature-based discovery should be regarded as a technique employed subsequent to text mining for explicit facts. General text-mining approaches focus on the extraction of relevant entities, for example genes and proteins, and relationships between them, example protein and protein interactions. The users of these approaches generally retrieve known, explicit co-occurrence-based knowledge that they are personally were not always aware of and text mining can be viewed as an efficient way of keeping abreast with the most important facts in the literature [8].

Building on large mined entities and facts LBD tools attempt to combine extracted information into truly novel hypotheses. Because many combinations of mined facts are possible, the main aim of LBD systems is to confine the explosively growing number of possible hypothesis to those that have the highest probability to be consistent. In contrast to text-mining and information extraction approaches, there is no straightforward evaluation possible as it is not easy to establish the correctness of generated hypotheses [8].

CHAPTER 3

SOFTWARE TOOLS AND TECHNIQUES

The necessity for techniques of text mining, information extraction (IE), and information retrieval (IR) has been increasing in various scientific fields so that appropriate information can be extracted from numerous scientific papers [13]. 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 and techniques 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 [3].

Other programs have enhanced discovery processes in science, such as [39, 17]. Some use combinatorial search as their basic approach, whereas others use more specialized methods that can exploit mathematical properties of the subject matter, such as strings in genomics. The aim here is not to provide a survey, but to state key concepts and illustrate them with a few programs that have led to published findings [18]. This chapter examines several published LBD systems, comparing their descriptions of domain and input data, techniques to locate important concepts from text, models of discovery, and experimental results. 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 chapter considers methods for automated systems based on data mining [3].

3.1. Medline Database

The MEDLINE [13] (Medical Literature Analysis and Retrieval System Online) database is a product of the US National Library of Medicine (NLM). Because of its coverage and free accessibility, MEDLINE is the most important database in field of biomedicine and is available for free searching on many websites of government and health agencies. It contains bibliographic citations and author abstracts from over 4,600 biomedical journals. Each citation is associated with a set of MeSH (Medical Subject Headings) that describe the content of the item. Presently the database comprises over 12 million records dating back to 1966. One of the most popular is the NLM Web based product PubMed [12]. MEDLINE is the primary component of PubMed [13].

3.2. Discovery Support Tools

As previously mentioned, literature-based discovery is distinctive from more generic text-mining approaches; however, generic text-mining applications might be used for LBD. Recently several systems have been developed specifically to support literature-based discovery and hypothesis generation (see table 1). A brief description is given of five of them that are freely available. An overview of some characteristics of these systems is given in Table 2. IRIDESCENT, the tool developed by Wren *et al.* has become commercial and is available from Texx Biopharmaceuticals, Inc. The ACS algorithms and viewer has only limited access. The Telemakus KnowledgeBase System, though freely available, will not be discusses because it has a much focused domain of application [8].

Arrowsmith, Manjal and BITOLA provide interfaces that enable user navigation through the generated results and corresponding literatures [16]. The use of an LBD tool will principally be more complicated than a straightforward literature search as there is no direct evidence of a generated hypothesis [8].

Arrowsmith/University of Chicago	http://kiwi.uchicago.edu/
Arrowsmith/University of Illinois at Chicago	http://arrowsmith.psych.uic.edu/
BITOLA	http://www.mf.uni-lj.di/bitola/
Manjal	http://suluinfo-science.uiowa.edu/Manjal.html/
LitLinker	http://litlinker.ischool.washington.edu/
ACS	http://www.biosemantics.org/
IRIDESCENT	http://www/etexxbio.com/
Telemakus	http://www.telemakus.net/

Figure 3.1: Currently available literature based discovery systems

3.2.1. Description of the LBD tools in Biomedicine

a. Arrowsmith/University of Chicago

Arrowsmith located at the University of Chicago is the original tool developed my Swanson and Smalheiser. The user has to upload two files that contain the results of two PubMed or OVID queries on A and C subjects, respectively. The server searches for overlapping title words and presents them to the user as the '*B*-List'. The user can edit the *B*-list and view the juxtaposed titles from both literatures for some selected *B*-terms. The user interface is rudimentary, and there is a steep learning curve. Currently Arrowsmith

can be used only for closed discoveries but the next version should also include the possibility of an open discovery.

Characteristics	Arrowsmith University of Chicago	Arrowsmith University of Illinois at Chicago	BITOLA	Manjal	LitLinker
Registration	No	No	No	Yes	Yes
Online/offline processing	Online	Online	Online	Offline	Online
Concept/terms	Title words	Title words + filtering of UMLS concepts in title words	MeSH and LocusLink	MeSH	UMLS
Documentation	Poor	Poor	Poor	Good	Average
Query formulation	PubMed or OVID (separated from tool)	PubMed (integrated in tool)	Term entry with feedback	Term entry without feedback	Term entry with feedback
Visualisation of results	List of terms, juxtaposition of titles	List of terms, juxtaposition of titles, linkout to PubMed	List of terms, linkout to PubMed	List of terms, linkout to PubMed	List of terms, indication of association strength, title and abstract
User interface	Poor	Average	Average	Average	Advanced
Application domain	General biomedicine	General biomedicine	General biomedicine + focus on genomics	General biomedicine	General biomedicine
Save sessions	Yes	Yes	No	Yes	Yes

Figure 3.2: LBD tools and their characteristics in methods in use

b. Arrowsmith/University of Illinois at Chicago

This tool is a re-implementation of original Arrowsmith at Smalheiser's lab at the University of Illinois at Chicago. Its major advances are a direct search in PubMed using PubMed's interface, semantic and frequency filtering of concepts, and a more polished user interface. Only closed discoveries are supported.

c. BITOLA

BITOLA uses an open discovery approach. The user starts with defining a query that is mapped to a concept X. Next, the user selects the category of interest, e.g. diseases, pathologic process. A rank-ordered list of relevant concepts Y that are directly

related to the query is then presented. Optionally, gene expression localizations can be selected. Next, one or more Y concepts are selected together with a target semantic category, e.g. a drug, and the result is a list of Z concepts that are potential discoveries. A linkout to PubMed with an AND query on the Y and Z concepts is provided to assist human assessment of the potential discovery.

d. Manjal

Manjal provides an open and a closed discovery option. Similar to BITOLA, the user has to select a semantic category of interest. Interestingly, when employing an open discovery process, the final results are automatically computed and without having the user to select the intermediate concepts. Computation takes some times. Therefore the results are not provided online. The user will receive an e-mail message when the results are available. Processing may take some minutes up to several hours, depending on the query and server load.

e. LitLinker

In LitLinker, an open discovery approach has been implemented. After defining a query, a list of resulting concepts is automatically generated without user intervention of selecting intermediate concepts. The user interface of presenting the results is highly informative and has been evaluated experimentally [8].

3.3. Scientific Discovery Programs outside Biomedicine

Grand challenges such as public health, security, genomics, environmental protection, education, and economics, are characterized by complexity, interdependence, globalization, and unpredictability. Although the unprecedented quantity of information surrounding these challenges can provide users with a new perspective on solutions, the

data surrounding complex systems vary with respect to levels of structure and authority, and include vastly different contexts and vocabularies. To be successful in this domain we must extend our models of information science such that they operate successfully in environments where the quantity of relevant information far exceeds our human processing capacity. For example, the well-accepted precision and recall metrics break down when hundreds of thousands of documents are relevant [31].

While LBD have biomedicine as the larger domain of research, there has been significant development of successful systems outside biomedicine. Two examples of successful systems-taken from mathematics, and chemistry that have enabled published discoveries in their respective literature as described in [17].

a. Graffiti

The Graffiti program developed at the University of Houston makes mathematical conjectures in such domains as graph theory and geometry (math.uh.edu/~siemion). Graffiti has motivated many graph theoreticians, including its designer, to try to refute or prove the generated conjectures which are broadcast on an email list. Many of the program's conjectures have been proven (by mathematicians) and published as regular mathematical contributions. Recent applications of Graffiti to chemistry have exploited the fact that molecules can be represented as graphs. The program keeps a database of previous conjectures so that when the program is run it does not repeat itself and instead will tend to produce novel conjectures [18].

b. Mechem

The Mechem program developed at Carnegie-Mellon University (in recent collaboration with A.V.Zeigarnik in Russia) [18] finds explanatory hypotheses (reaction

mechanisms) in chemistry. That is, given the starting materials of a chemical reaction, any observed products and intermediates, and prior background knowledge expressed as constraints, the program finds all simplest mechanistic hypotheses that explain how the products are formed while respecting the constraints [17]. The program's mechanisms tend to contain novelty because the pieces (elementary reactions and chemical substances) that make up a hypothesis are not drawn from any stored catalogue of common reactions; rather, they are generated from basic principles using algorithms minimally slanted toward particular solutions.

The mechanisms are often interesting because are the simplest, that is, the program reports mechanisms that contain fewest intermediate substances and steps. The mechanisms are understandable because the space being searched is taken directly from chemistry. Finally, the output is plausible because the user articulates and objections, via a graphical interface that allows for well over 100 kinds of constraints, and runs the program again with augmented input. This interaction repeats until no further problems remain [18].

3.4. Swanson and Smalheiser (1997) Arrowsmith System

Arrowsmith, created by Don Swanson and refined with the help of his collaborator Neil Smallheiser, performs hypothesis discovery by an algorithm that is astonishing in its simplicity, which is best described by a scenario. Suppose a patient who is receiving a recently introduced drug is found to have a rare adverse event. The question is whether the agent, rather than the underlying disease, could have been responsible. [21]. Arrowsmith is a unique computer-assisted strategy designed to assist investigators in

detecting biologically-relevant connections between two separate sets of articles in Medline [20].

3.4.1. Why do we need Arrowsmith System?

Arrowsmith is basically a research tool for studying complementary noninteractive structures in the scientific literature and at the same time to create a working system useful to biomedical scientists [15]. It has been able to produce surprising results using very simple methods of data access [22]. Arrowsmith is arguably the best established system for carrying out data mining of the biomedical literature, having been widely analyzed, replicated and discussed by the information science community [20]. Scientists routinely make use of the same information-gathering strategy in deciding which experiments to pursue next, but ARROWSMITH facilitates this task by automatically and systematically creating a comprehensive list of B-terms to be considered [19].

3.4.2. Arrowsmith Algorithm

Simple Semantic Match

i. a. Collect all publications that contain term A. This will be literature A.

b. Collect all publications that contain term C. This will be literature C.

ii. Find all concepts that are in the A list and in the C list. There is no advanced matchmaking here; matches are made only between identical URI's. The resulting lists of matched concepts are the "bridging" terms, and they make up the B list.

iii. Return a list of the papers that connect A to B, and B to C [22].

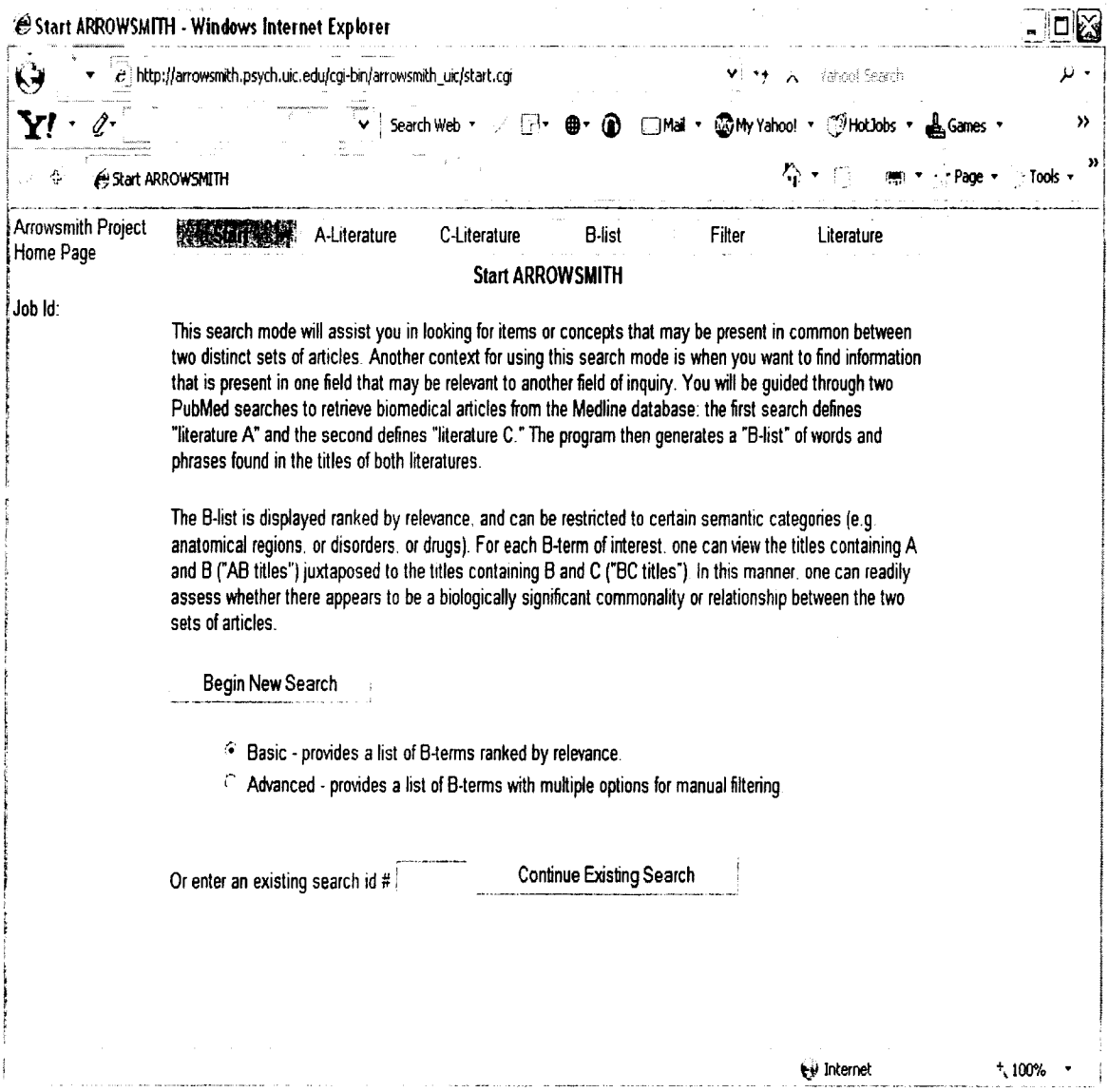

Figure 3.3: Arrowsmith System/University of Illinois

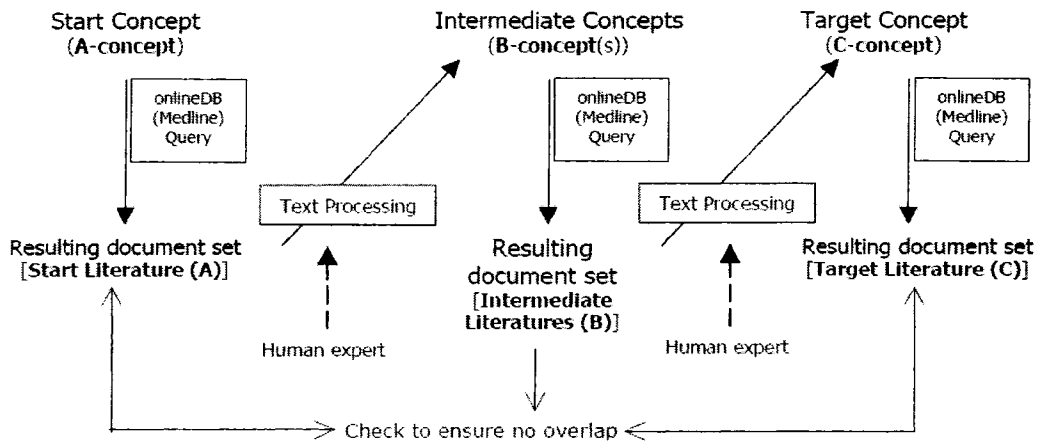

Figure 3.4: Swanson's Framework for Arrowsmith system

3.4.3. Arrowsmith – Materials and Methods

The web-based version of Arrowsmith uses a closed model of discovery [3] and it employs the two node search technique [20]. 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*) [3]. We begin at the point where an investigator seeks to assess the possible relationships that may exist between two items, A and C, which may represent the names of physiologic parameters (e.g. blood pressure), diseases, drugs, nutrients, proteins, etc.

A relation between A and C may already be established by experimental, epidemiologic or genetic linkage studies, or the user may simply wish to hypothesize that A and C are related in some manner. Two MEDLINE searches are carried out by the user, creating an A file consisting of all titles containing the term A (and synonyms and alternative spellings, if appropriate), and a C file consisting of all titles containing the term C. Then the ARROWSMITH software simply compiles a list of all words and phrases B common to the two sets of titles. Many of the B-terms will be predictably non-

interesting (e.g. 'the,' 'patient') and are excluded using a pre-compiled stop-list of 5000 words. One can also apply a more stringent criterion, if desired, so that items will appear on the B-list only if they appear in two or more titles in both the A and C sets.

The investigator then scrutinizes the B-list, looking for items that he or she feels might plausibly link A and C. For each such B-term that is identified, another pair of lists is compiled, to obtain an AB list of titles that contain both the terms A and B, and a BC list of titles that contain both B and C. By juxtaposing these lists, the investigator can quickly judge whether the titles indicate a biologically significant relationship linking A, B and C.

Such a relationship may, in fact, be well-known in the field, even if it was not known by the investigator doing the search; in such cases, ARROWSMITH may be regarded as providing a type of information retrieval that extends the basic capability of MEDLINE searches [19]. Figure 3.6 depicts the current Arrowsmith system developed at the University of Illinois. The system operates in several stages and supports closed discovery process.

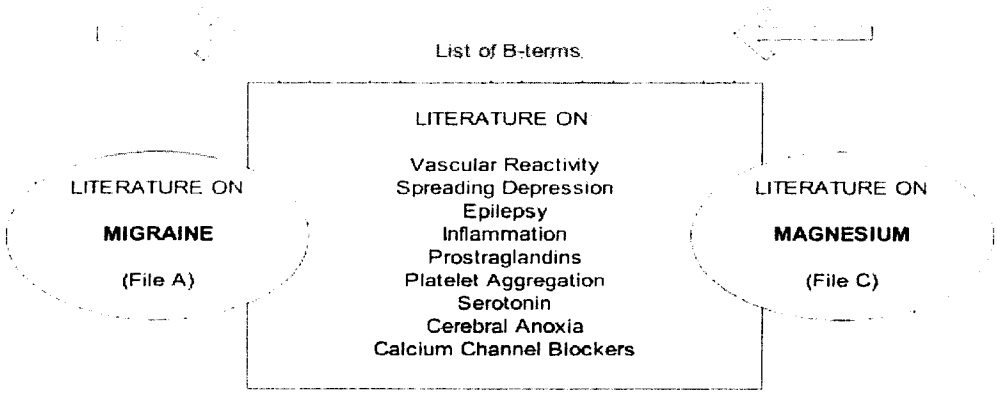

Figure 3.5: Venn diagram of connections between 2 titles

3.4.4. Results and Discussion

The wide range of potential applications of ARROWSMITH can be appreciated by giving two specific examples. Figure 3.5 shows an intersection of two articles that have common phrases and words.

3.4.4.1. Suggesting new therapeutic approaches

An early ARROWSMITH search explored possible ways in which Magnesium (Mg) may be linked to Migraine. At the time this analysis was performed, no prior papers in the literature had tested whether there was a biologically-significant relationship between Mg and migraine. Yet 11 B-terms existed whose AB and BC titles were each strongly suggestive of such a relationship. Moreover, these linkages were consistent: one would expect a local or systemic deficiency of Mg to worsen migraines, and Mg supplementation to prevent or ameliorate migraines.

Dietary fish oil was linked to Raynaud's disease in a similar fashion. Multiple B-terms indicated that a set of physiologic parameters were well-known to be abnormal in Raynaud's disease, and dietary fish oil was well-known to alter the same parameters in a direction which should normalize the abnormalities in the disease. Yet at the time the ARROWSMITH search was performed, no papers had examined whether dietary fish oil would ameliorate the signs and symptoms of Raynaud's disease [19].

3.4.4.2. Anticipating adverse drug reactions

For example, anti-inflammatory agents such as indomethacin are currently being examined for their utility in treating or preventing Alzheimer's disease. Though the primary action of indomethacin is clear (inhibiting prostaglandin synthesis), indomethacin has so many effects in so many organs that it is difficult to anticipate

which, if any, effects of indomethacin might be expected to be of special relevance to patients with Alzheimer's disease. Researchers have carried out an ARROWSMITH analysis with A=indomethacin and C=Alzheimer's disease, looking for terms B that might plausibly link these items. Since Alzheimer's patients are thought to have decreased cholinergic activity within the cerebral cortex that contributes to their dementia, this may be regarded as a potential adverse drug reaction, and those involved in clinical trials with indomethacin should at least be alert to this possibility.

3.5.Issues in Arrowsmith System

Arrowsmith provides interfaces that enable user navigation through the generated results and corresponding literatures. However, it is limited in the ways researchers can navigate and evaluate their results. For example, with these interfaces, it is not possible to retrieve all the target terms connected to a selected linking term or all the linking terms that connect a selected target term to the starting term. Such multi-dimensional navigation capabilities could prove essential for researchers who want to evaluate the quality and validity of the potential connections that the system generates [16]. Arrowsmith has been able to produce surprising results using very simple methods of data access. It is clear that their method of searching only on title keywords will inevitably miss some connections and have problems detecting direct relationships between terms. An extension of the project to the semantic web is a natural one, allowing the algorithmic method to reach new depths in the literature [4].

Although Arrowsmith system is operated as an experimental prototype, it may eventually become widely available as a supplement to current database search techniques [27]. Arrowsmith's conjectures tend to be novel because the methodology

involves a citation analysis to verify that no (or few) MEDLINE articles cite both subliteratures responsible for the associations AB and BC, so that there is no evidence from MEDLINE that anyone has noticed these connections. Swanson's colorful term for the program's output is "Undiscovered public knowledge" [17].

3.6. Existing LBD Algorithms and Methods

Following Swanson's Model in LBD, several researchers have formulated their methods which employ various text mining techniques and Aprirori and Hoffman algorithms. We have discussed a few popular techniques by different investigators.

3.6.1. Gordon and Lindsay – Information Retrieval Techniques

Gordon and Lindsay (1996) also attempt to replicate Swanson's model of discovery. They use techniques form Information Retrieval including token frequency (a metric summing the occurrence of tokens such as words), record frequency (number of records (e.g., documents) containing a given token) and tf*igf (token frequency * inverse global record frequency, a traditional IR technique). Their idea is to use the correlated statistics to identify intermediate literature with strong conceptual similarity to the starting point [3].

3.6.2. Gordon and Dumais – Latent Semantic Indexing

Gordon & Dumais (1998) offer an alternative method to support LBD. They use Latent Semantic Indexing (LSI), which employs latent semantics based on higher-order-co-occurrence to compute document and term similarity. LSI can reveal hidden relationships among terms, as terms semantically similar lie closer to each other in 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 [3]

3.6.3. Gordon and Lindsay – Trigrams and Contextual Analysis

Gordon and Lindsay (1999) proposed an analytic approach based on the word frequency statistics used in information retrieval research [30]. 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. 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 [3].

3.6.4. Weeber et al. – 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) [3].

3.6.5. Pratt and Yetisgen-Yildiz – 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 [40]. This is one of the few LBD studies that

employ 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 10 ARM was applied to LBD previously by Hristovski, Stare, Peterlin and Dzeroski (2001). 20 differences being that in ARM the tri-occurrence, quad-occurrence, etc. of terms are discovered [3]. 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) [40].

3.6.6. Van der Eijk et al. - Associative Concept Spaces

Van der Eijk et al. (2004) propose a novel algorithm for finding associations between related concepts present in literature. The user output is a visual graph displaying closeness of concepts instead of a ranked list, the classic A->B->C approach is not directly followed, and concepts are analyzed using a multi-dimensional [3] space by a hebbian type of learning algorithm [23]. Co-occurrence is the central concept of this approach. From the visual output authors conclude 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 [3].

3.7. Discussion of Automation of the LBD Process

Text data mining should be useful for anticipating new technologies and new uses for existing technologies, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature [44]. 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). There are many domain, data, and goal- specific challenges in fully automating LBD systems.

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 [3].

Although Arrowsmith software was employed to help define and juxtapose the two literatures in question, these advanced programs were not essential, and indeed we found in this case that the same outcome was obtained using conventional Medline searching techniques alone. The critical factor was the overall strategy of approaching the problem:

first, to define two specific fields explicitly that are hypothesized to contain complementary information; second, to identify common factors that bridge the two disciplines and third, to progressively shape the query once initial findings are obtained. Thus, in contrast to some current perceptions, the process of text data mining is neither automatic nor is it restricted to those who have access to customized computer systems [44].

CHAPTER 4

IMPLEMENTATION OF LBD TOOL

In order to observe and analyze the performance measures of the available literature based discovery systems and to model its functional parameters, a software program was coded, as a part of this thesis. We use the general architecture of the Arrowsmith system developed by Swanson and Smalheiser (1997) as our guideline and focus on replicating Swanson's model of LBD process. It is an effort to aid the analysis of various LBD system configurations. The system is based on a three-way interaction between computer software, a bibliographic database (such as Medline), and a human operator.

4.1. Experimental Setup

This chapter describes the functions available in the tool and flexibility it provides in testing the LBD methods. The software was implemented in Ruby programming language with a front end in Java Swing. The minimum requirements for the software tool to run are a computer system with optimum configuration (300Mz Processor with 128MB RAM) that is running on Windows, UNIX or Linux, Java Runtime Environment installed in the computer, preferably with J2SDE 1.4.2 and RUBY compiler of version 182-21.

The software tool is opened by double-clicking the executable file, which contains the generated class files of the source code. The LBD tool is user friendly and enables the user to perform experiments in simple step by step execution. The goal of the work

reported here is to evaluate Arrowsmith. The computer is used to search for, organize, and display information for users, who then look for implicit connections that may suggest novel, plausible scientific hypotheses [9].

4.2. Project Description

The Literature based discovery tool that is developed is a replication of Don Swanson's Arrowsmith system. The Closed model of Swanson's discovery process described in chapter 2 is implemented [25]. Initially a problem is defined to carry out the experiment in the LBD tool. The tool is programmed to operate in several stages. The Ruby programming was used to implement basic Natural Language Processing techniques (NLP) on the documents such as stop word elimination and stemming algorithms.

We begin at the point where an investigator seeks to assess the possible relationships that may exist between two items, A and C. A relation between A and C may already be established by experimental, or the user may simply wish to hypothesize that A and C are related in some manner. Two database searches are carried out by the user, creating an A file consisting of all titles containing the term A (and synonyms and alternative spellings, if appropriate), and a C file consisting of all titles containing the term C. Then the Literature based discovery software simply compiles a list of all words and phrases B common to the two sets of titles. Many of the B-terms will be predictably non-interesting (e.g. 'the,' 'test') and are excluded using a pre-compiled stop-list words [19]. We also implement the Porter Stemming algorithm on the remaining list of B-terms to stem the words to avoid occurrences of multiple terms.

The investigator then scrutinizes the B-list, looking for items that he or she feels might plausibly link A and C. For each such B-term that is identified, another pair of lists is compiled, to obtain an AB list of titles that contain both the terms A and B, and a BC list of titles that contain both B and C. By juxtaposing these lists, the investigator can quickly judge whether the titles indicate a biologically significant relationship linking A, B and C. Such a relationship may, in fact, be well-known in the field, even if it was not known by the investigator doing the search; in such cases, the LBD tool may be regarded as providing a type of information retrieval that extends the basic capability of database search strategies [19].

4.3. Flowchart Model of LBD tool

Figure 4.1 is a flowchart illustration of the LBD tool's working model. The procedure begins with a downloading of titles from literatures A and C, and proceeds to the lower right where an output display is produced as a heuristic aid for the human user of the system. The display is organized to facilitate comparison of A-titles with C-titles for each B-term that they have in common, and serves as a guide to the literature [37].

4. 4. Project's Main Window

The figure 4.2 shows the main application frame that opens up, as soon as the software is opened. The application looks like any other window application with similar look-and-feel and is easy to navigate. It has a Menu bar that has got File, Results and Exit menu options.

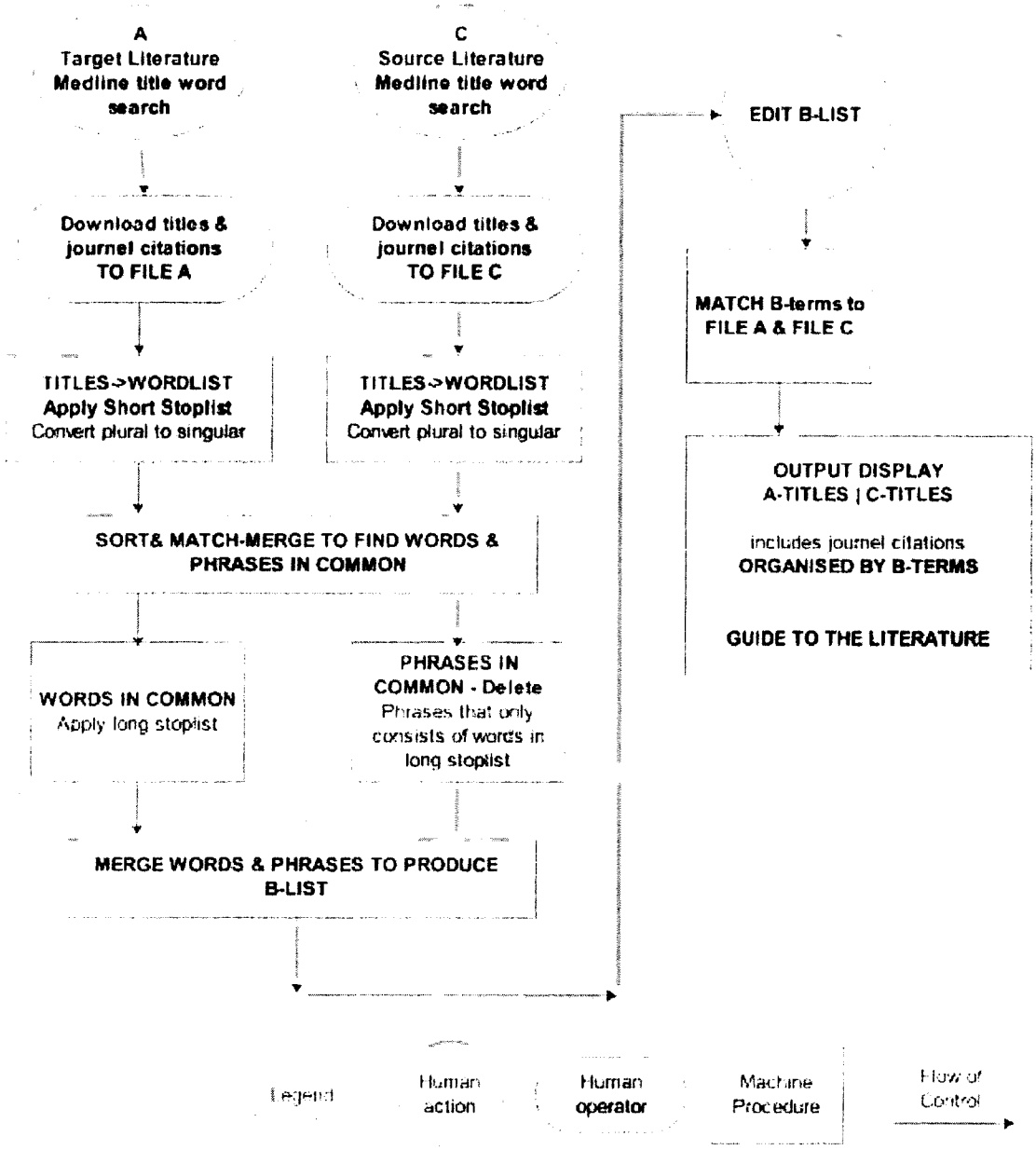

Figure 4.1: Flowchart description of the LBD tool and various stages

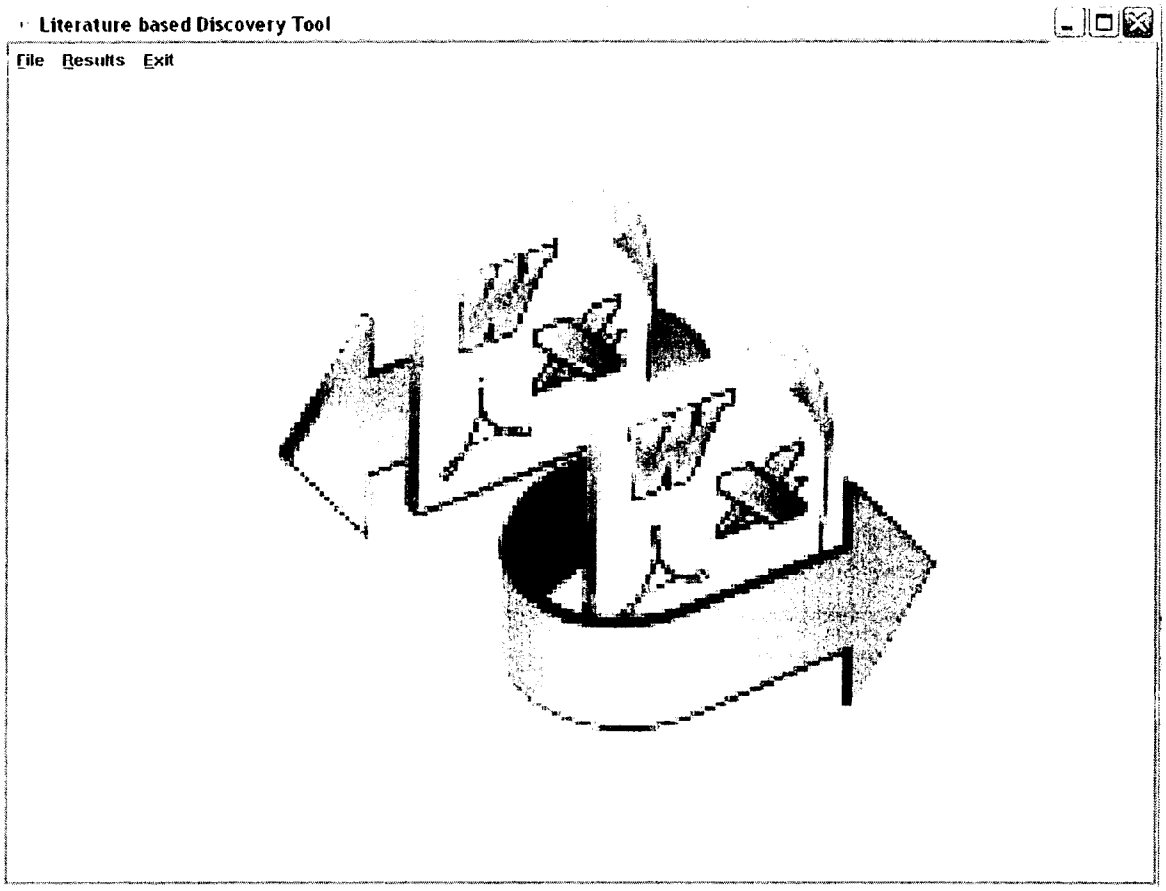

Figure 4.2: Literature based discovery tool's Main window

File menu has three menu items, namely, Open which opens up the simulation run window, Close which closes the active window, and Exit which terminates the application. The Results Menu again has two menu items, View Results opens the View Results window and Export Results exports the simulation run information to a custom text editor where it can be viewed, saved or printed. The Exit menu terminates the application.

4.5. Operation Stages of the LBD tool

4.5.1. Implications of A-C intersection

The basic step to conduct any experiment is to define the problem and set of evaluation procedures. The goal of discovery is to find knowledge that is novel, interesting, plausible, and understandable [17].

We first start with framing a concept of interest, which we call it the source title and denote it as concept C. Then we analyze the possible target for the source that we have defined. The target is denoted as concept A. We ensure that articles belonging to Concept A and Concept C do not co-cite each other. It is necessary to understand the implications of A-C intersection for conducting the discovery process. Initial Procedure should always be preceded by a database search for all records that contain both A and C (not restricting the search to just titles), in order to identify any A-C or A-B-C relationships that are already explicitly published [26].

- ✓ Disjoint literatures: Best opportunity for finding previously unknown connections.
- ✓ Small overlap: May be as good, or better.
- ✓ Large overlap: not promising for novelty.

Such known B-linkages should be investigated in advance of executing the whole experiment to avoid unknowingly rediscovering them as indirect linkages; further experiment process can then focus on relationships that are either novel or at least cannot be found by conventional searching. Understanding strategies used in conventional searching is important at this point; finding the "direct" literature is not always straightforward. The citation interaction pattern also can play a key role in determining whether an A-C relationship exists and is already known, a process described in more

detail elsewhere. Thus, a discovery program that too often leads to familiar, dull, wrong, or obscure knowledge won't be used [18].

4.5.2. Creation of Files

After we have defined our problem we then proceed to execute it using the LBD tool. First conduct a subject title-word search for the word or term denoted by concept C, and downloads all records found and save it as File C. Next, create a similar but separate computer file based on the title-word or term corresponding to concept A and save it as File A. Title-word searching may be enhanced by including subject-headings as well [33]. No title contained both words [15]. A focus on searching title words and phrases may be important, for three reasons:

a. It tends to improve precision

b. It improves the odds that B-terms are meaningfully linked to their corresponding A and C terms, because they are closer to them in titles than in abstracts.

c. Complementarity is easier to recognize and the amount of text to be examined is much less than in scanning through complete records [26].

4.5.3. Stage 1- Uploading File A and File C

The user clicks the upload button on the File Menu. It prompts the user to upload to the first file to the tool from the location it is saved. File A is uploaded. The same procedure is carried out another time and File C is uploaded. The files are then transmitted (uploaded) one at a time. After uploading both File A and File C, the user now clicks the button named Process.

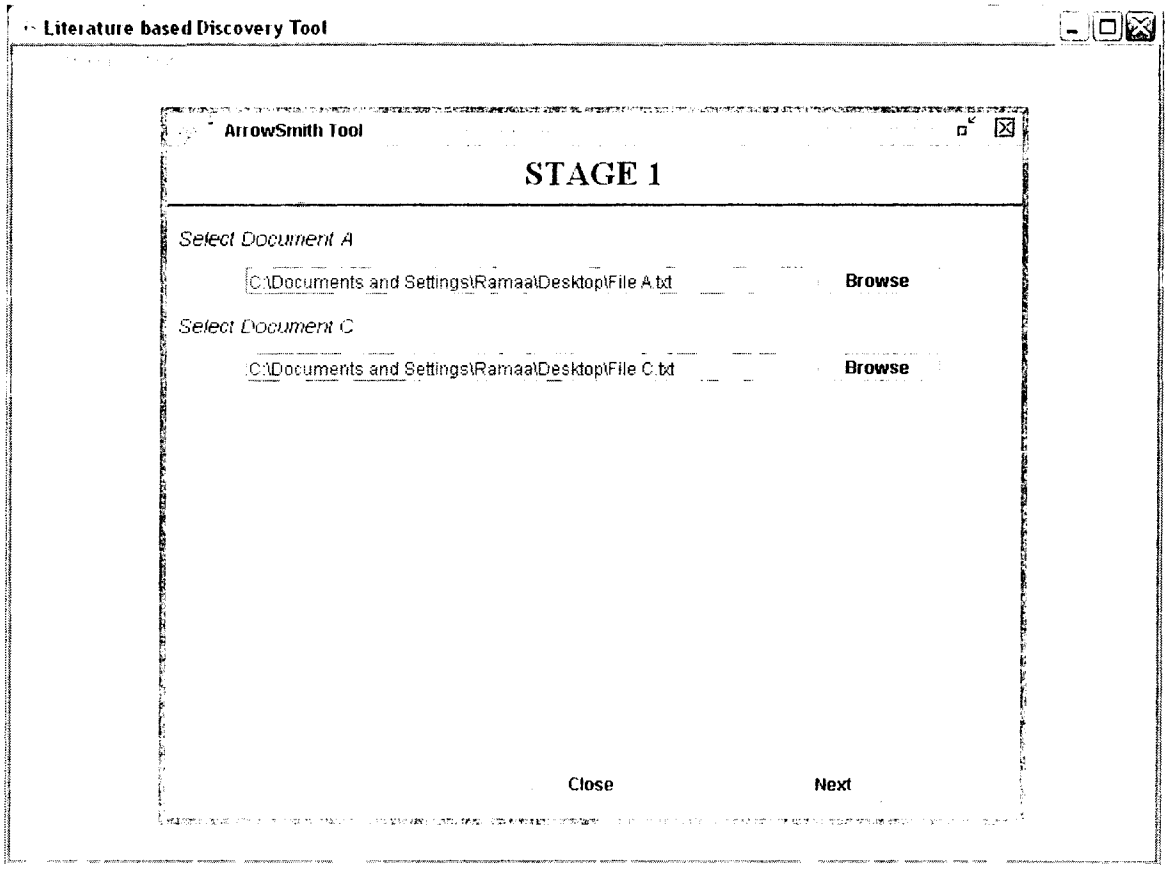

Figure 4.3: Stage 1 – Uploading File A and File C

4.5.4. Stage 2 – Creation of B-list

After uploading File A and File C to the tool, the RUBY programming performs operations such as stopword list elimination and stemming the words and phrases in both the literatures. The program then generates the B-list, which is the intersection of words and phrases common to literatures in File A and File C. If the two input sets each consist of thousands of records, the resulting number of key B-terms they have in common may be so great as to impede a careful search for the few that are novel and of scientific interest. We address this problem on two fronts: first, by trying to improve the search

strategies used in creating Files A and C, and second, by filtering and organizing the B-list itself [9].

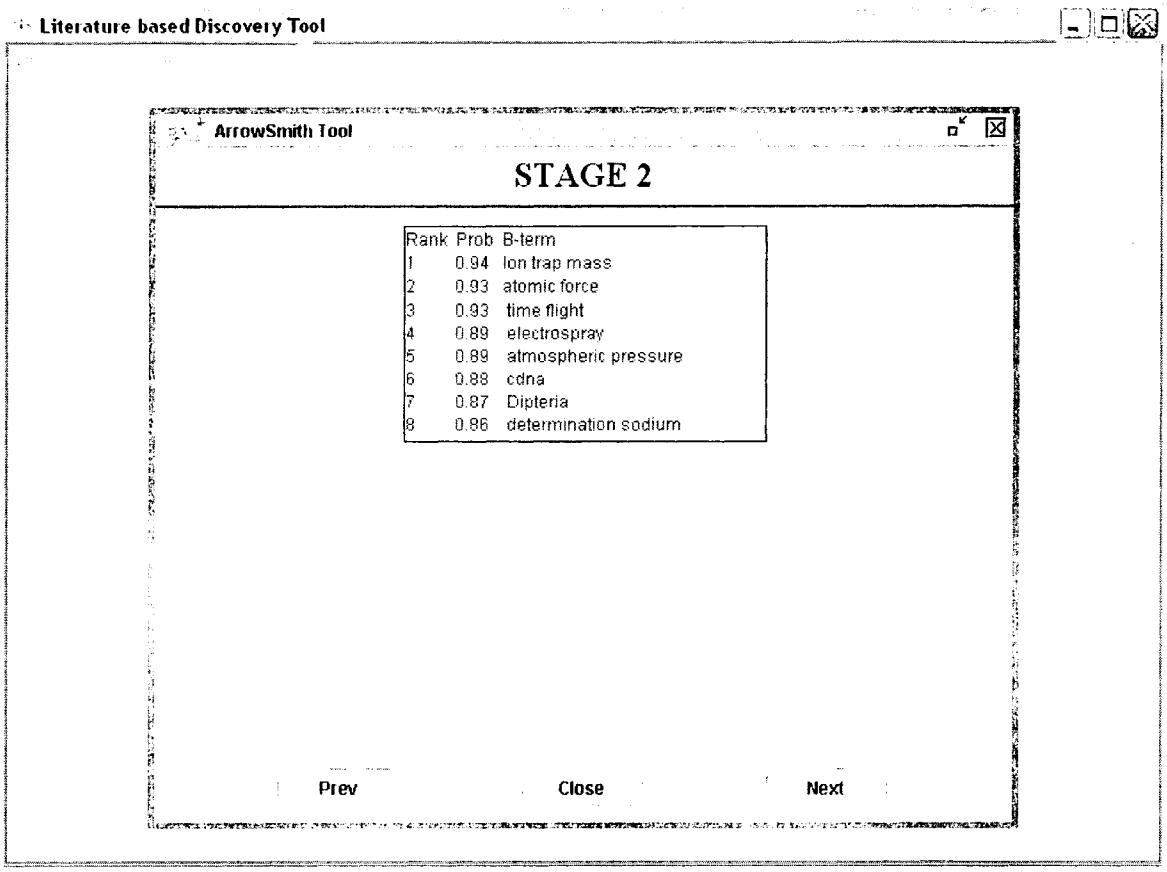

Figure 4.4: Stage 3 – Creation of B-list with Ranking & Probability

4.5.4.1. Stopword List Elimination

Stoplists, lists of words to be excluded because they are predictably of no interest, are often used in searching text. The 10 most frequently occurring words in English accounts for 20-30% of all the tokens in a document. Most stoplists consist of up to a few hundred words that are not subject-oriented. A disadvantage of a very long stoplist is that if it does cause useful terms to be omitted from the B-list, the user is unaware of the loss.

We continue to study the consequences of different types and lengths of stoplists. When using the term weighting technique, document representation is done by first removing functional words (e.g. conjunctions, prepositions, pronouns, adverbs, etc.) [28]. Any stoplist, once compiled, becomes part of the machine procedure; it is always to some extent fallible, but open to inspection, criticism, and improvement [9].

4.5.4.2. Stemming of words and phrases

Stemming is employed, which collapses singular and plural variants of terms [3]. We have implemented stemming process using Porter stemming algorithm. The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems [34]. Frequently, the performance of an Information Retrieval system will be improved if term groups such as this are conflated into a single term. In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous [35].

4.5.5. Editing the B-List

Any *B*-term that is judged by the user to be of scientific interest because of its relationship to the *A* and *C* literatures is called a "target." It is the target terms that potentially may lead to literature-based discovery. Typically, only a few targets are found among hundreds of B-terms examined [9]. The *B*-LIST is next edited by removing redundancies and nonuseful terms. The editing process in general may include adding,

deleting, removing redundancies, or revising entries in order to compensate for certain limitations in the mechanized rules.

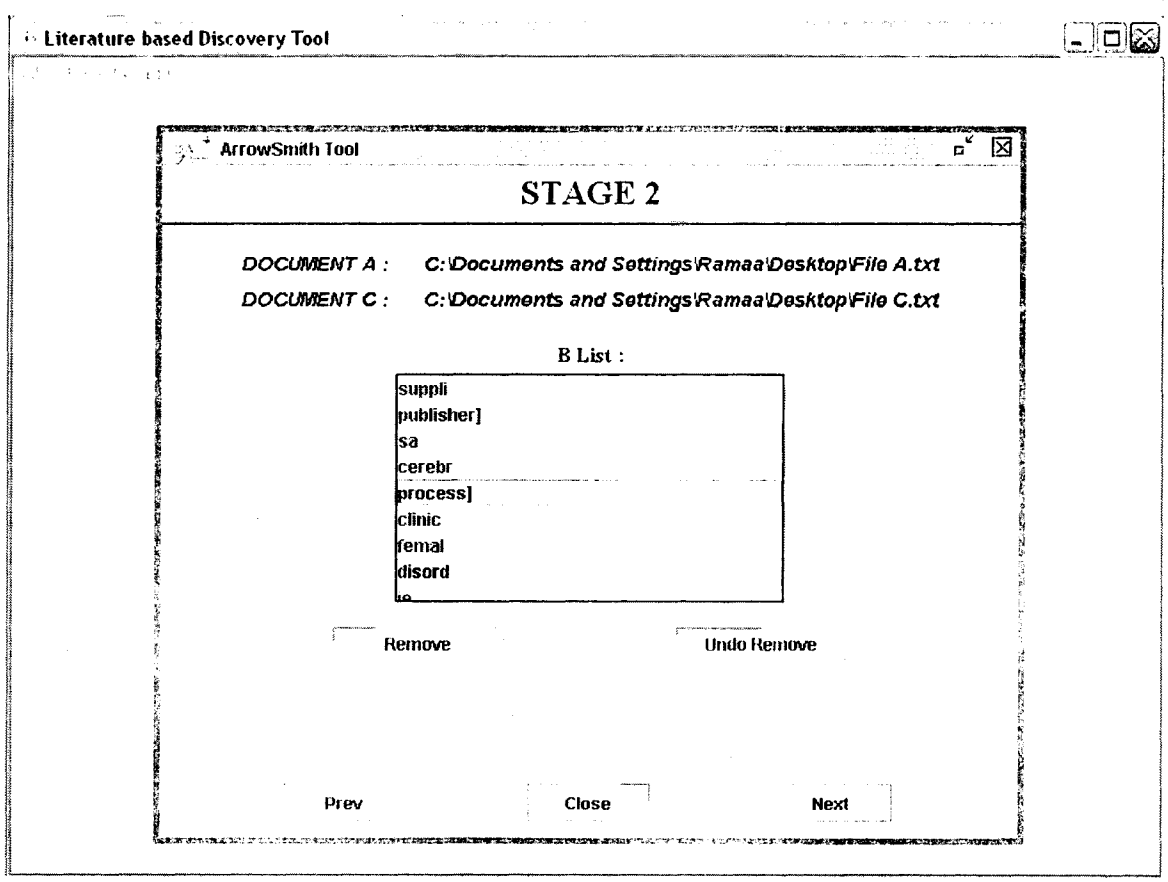

Figure 4.5: Manual Deletion of unwanted B-terms from the B-list

4.5.6. Stage 3 - Final Output Display and title Browsing

After the editing has been completed, the computer produces a display or printed output that shows titles within title file *A* that share *B*-terms with titles in file *C*, organized alphabetically by the *B*-terms. A similar display is created for title file *C*, thus facilitating a comparison of titles *A* with titles *C* that have one or more *B*-words or phrases in

common. The display is organized to facilitate comparison of *A*-titles with C-titles for each *B*-term that they have in common, and serves as a guide to the literature [15].

AB literature	B-term	BC literature
("citrus"[TIAB] NOT Medline[SB...	atmospheric pressure	glow[All Fields]
1: Mode of action of a novel nonchemical method of insect control: atmospheric pressure plasma discharge. 2006 Add to clipboard		1: Surface oxidation of polyethylene using an atmospheric pressure glow discharge with liquid electrolyte cathode. 2006 Add to clipboard
2: Multiple-stage mass spectrometric analysis of six pesticides in oranges by liquid chromatography-atmospheric pressure chemical ionization-ion trap mass spectrometry. 2004 Add to clipboard		2: [Experiment on optical radiation characteristic of low temperature plasma at atmospheric pressure] 2006 Add to clipboard
3: Analysis of bitter limonoids in citrus juices by atmospheric pressure chemical ionization and electrospray ionization liquid chromatography-mass spectrometry. 2003 Add to clipboard		3: An experimental study of the electrospraying of water in air at atmospheric pressure. 2004 Add to clipboard
4: Analysis of plant sterol and stanol esters in cholesterol-lowering spreads and beverages using high-performance liquid chromatography-atmospheric pressure chemical ionization-mass spectroscopy. 2003 Add to clipboard		4: Biocompatibility evaluation of ePTFE membrane modified with PEG in atmospheric pressure glow discharge. 2002 Add to clipboard
5: Liquid chromatography/atmospheric pressure chemical ionization-mass spectrometric analysis of benzoylurea insecticides in citrus fruits. 2000 Add to clipboard		5: An atmospheric pressure glow discharge optical emission source for the direct sampling of liquid media. 2001 Add to clipboard
		6: Liquid sample injection using an atmospheric pressure direct current glow discharge ionization source. 1992 Add to clipboard

Figure 4.6: Output display – Intersection of AB and BC terms of a common B-term

Figure 4.6 is a sample of the final output display of files. Basically, when the user clicks on a particular B-term in stage 4 to view its corresponding literature, the intersection of AB and BC literatures for the specific B-term that is being clicked is displayed. This is probably the most interesting part in the construction of the LBD tool.

The implementation of stage 5 will serve as a key factor for any researcher to find maximum connections between the A-C literatures via the common B-term. As mentioned in Swanson's ABC model in evaluating closed discovery methodology, if there exists maximum links between the literatures, there is maximum possibility of making a hidden connection between concept A and concept C. We can also see that both the articles in the AB and BC literature do not co-cite each other or have A or C concepts being mentioned in either articles.

4.6. Analysis of the Literature Based Discovery Tool

Scanning or browsing selected lists of titles that have key words n common is helpful as an initial source of clues to possible casual connections [37].Researchers can use this method to mine the large body of scientific literature, which is increasing at an exponential rate. The predicted new relationships can serve as candidates for new research themes, as impetus for inspiration, or as hypotheses to be tested in future.

Based on their research interests and background knowledge, researchers could choose some of the new relationships as future research directions. This method can also be applied to other disciplines, if a controlled vocabulary is set up to index and search the literature in database, and if the controlled vocabulary is arranged in a hierarchical structure. Actually, these two requirements are the system engineering results of acquiring domain knowledge from discipline experts [30].

Also, the LBD tool developed has been incorporated with a technique that helps the user to export the results into two separate files. This feature helps the user to save specific search results in source and target files for which he would wish to perform

further analysis. Figure 4.7 is a sample of a result called BC Hypothesis – Result that is being exported to a separate text file.

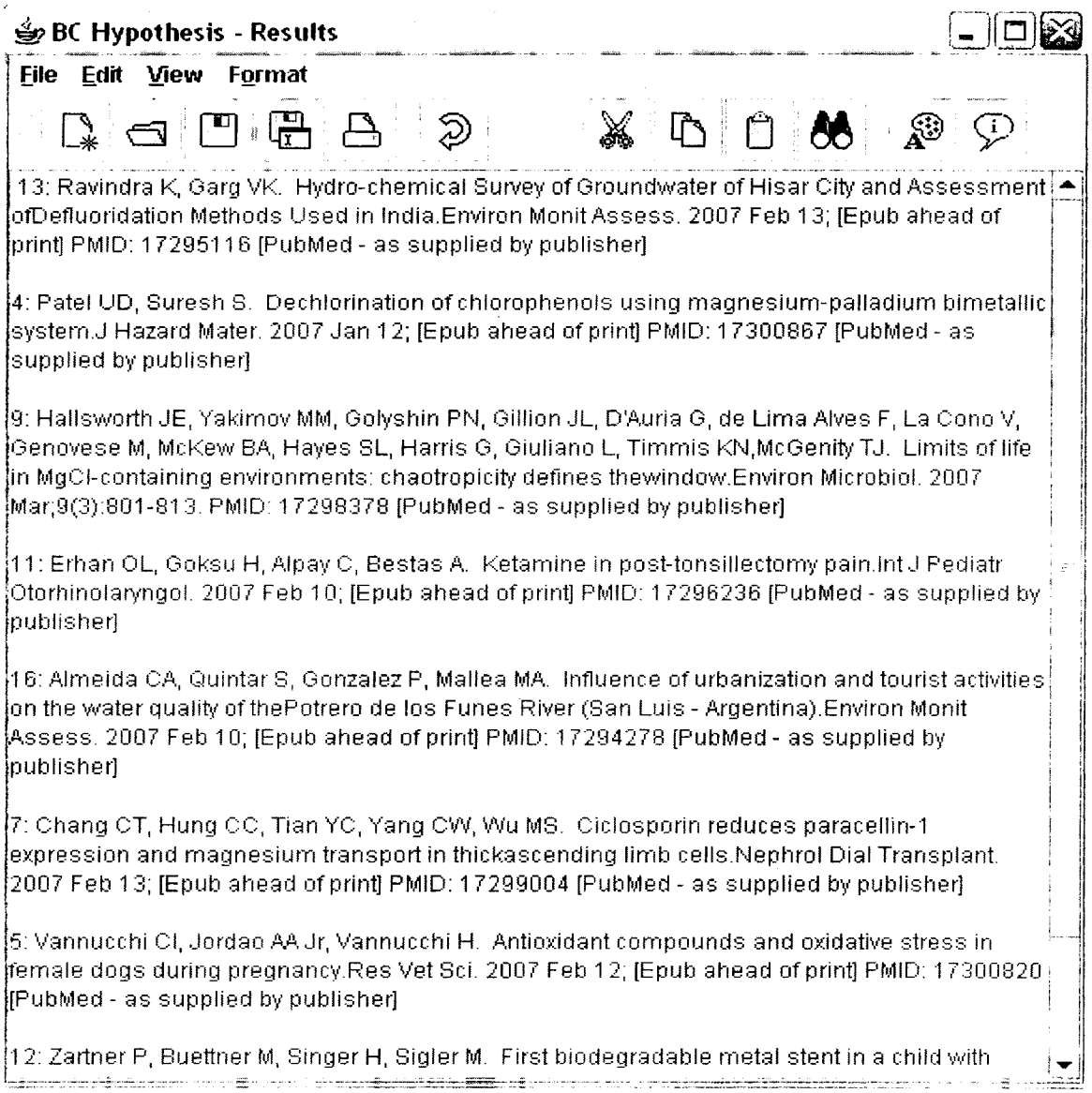

Figure 4.7: Exporting results into separate files for literature analysis

CHAPTER 5

EXPERIMENTAL EVALUATION

5.1. Experimental Datasets

We chose to test the Migraine-Magnesium Hypothesis as part of the experimental evaluation of the LBD tool. For example, when investigating causes of migraine headaches, we extracted various pieces of evidence from titles of articles in the biomedical literature. Some of these clues can be paraphrased as follows:

- Stress is associated with migraines
- Stress can lead to loss of magnesium
- Calcium channel blockers prevent some migraines
- Magnesium is a natural calcium channel blocker
- Spreading cortical depression (SCD) is implicated in some migraines
- High levels of magnesium inhibit SCD
- Migraine patients have high platelet agreegability
- Magnesium can suppress platelet aggregability

These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache; a hypothesis which did not exist in the literature at the time Swanson found these links [7]. Hereby we consider two datasets consisting of all the titles for a search independently on Migraine and Magnesium with Migraine as File A and Magnesium as File C.

(a) Statements supported by migraine literature	(b) Statements supported by magnesium literature
1a—Stress and Type A behavior are associated with migraine.	1b—Stress and Type A behavior lead to body loss of magnesium.
2a—Excessive vascular tone and reactivity can aggravate or predispose to migraine.	2b—Magnesium can reduce vascular tone and reactivity.
3a—Calcium channel blockers have been used to prevent migraine.	3b—Magnesium is a natural calcium channel blocker.
4a—Spreading cortical depression may be implicated in the early phase of a migraine attack.	4b—High levels of magnesium in the extracellular cerebral fluid can inhibit spreading cortical depression in animals.
5a—There is evidence for a connection between epilepsy and migraine.	5b—Magnesium deficiency can increase susceptibility to epilepsy in animals.
6a—Migraine patients have abnormally high platelet aggregability.	6b—Magnesium can inhibit platelet aggregation.
7a—Platelets from migraine patients are abnormally sensitive to serotonin release.	7b—Magnesium deficits can lead to high levels of serotonin release.
8a—Substance P may be a cause of head pain in migraine.	8b—Magnesium deficits can increase Substance P activity.
9a—Low levels of prostacyclin or high prostaglandin e1 release can aggravate vasoactivity and Substance P activity in migraine.	9b—Magnesium deficits can lead to low levels of prostacyclin release.
10a—Migraine may involve sterile inflammation of cerebral blood vessels.	10b—Magnesium has anti-inflammatory properties.
11a—Cerebral hypoxia may play a key role in migraine.	11b—Magnesium can protect against brain damage from hypoxia.

Figure 5.1: Eleven indirect arguments connecting magnesium and migraine

5.2. Experiments

5.2.1. Search for File A and File C

We then downloaded the articles that has Migraine and Magnesium as its title word and saved them separately in two text files. The next two figures show the samples of File A and File C.

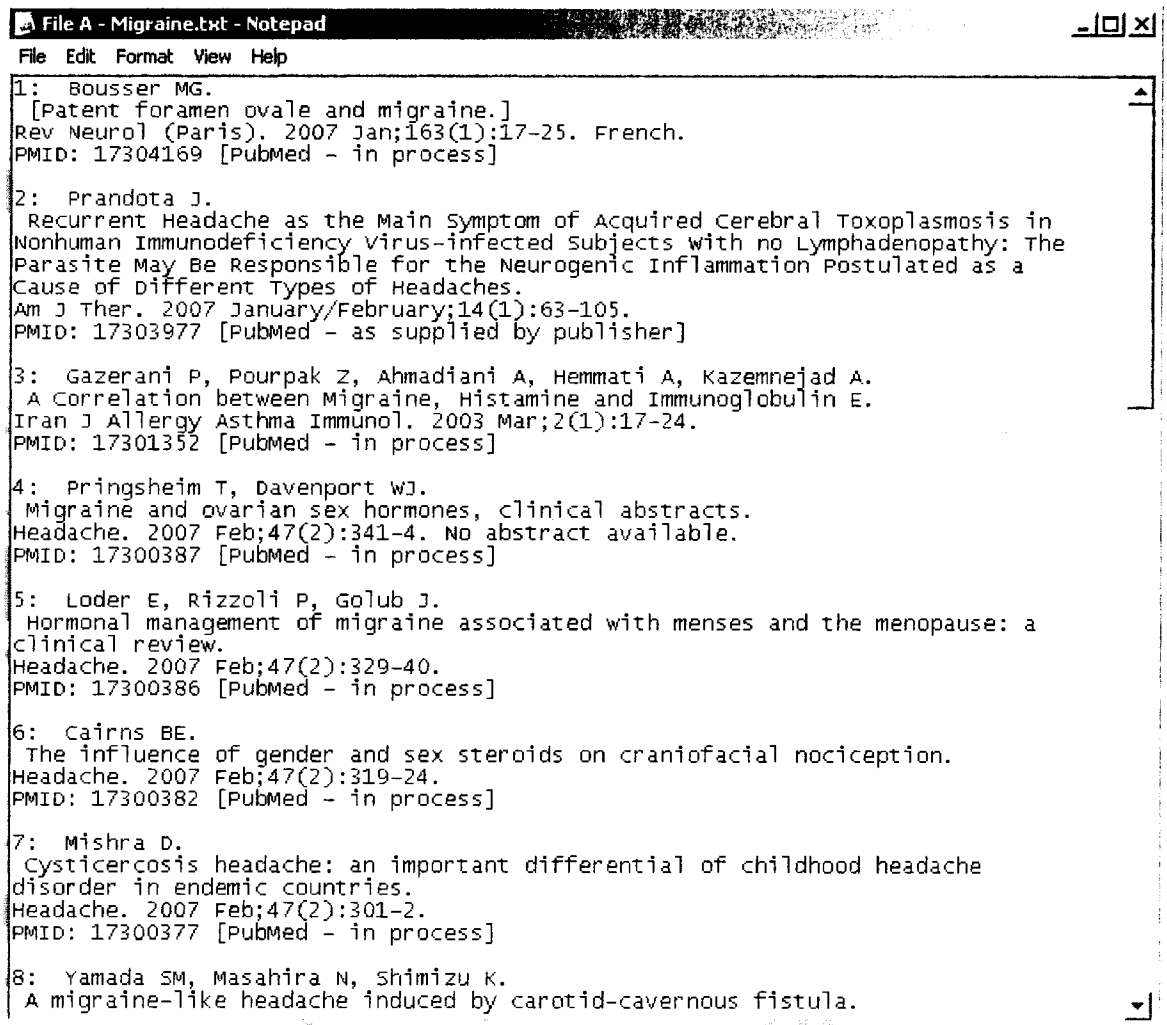

Figure 5.2: Sample display of File A articles of title Migraine

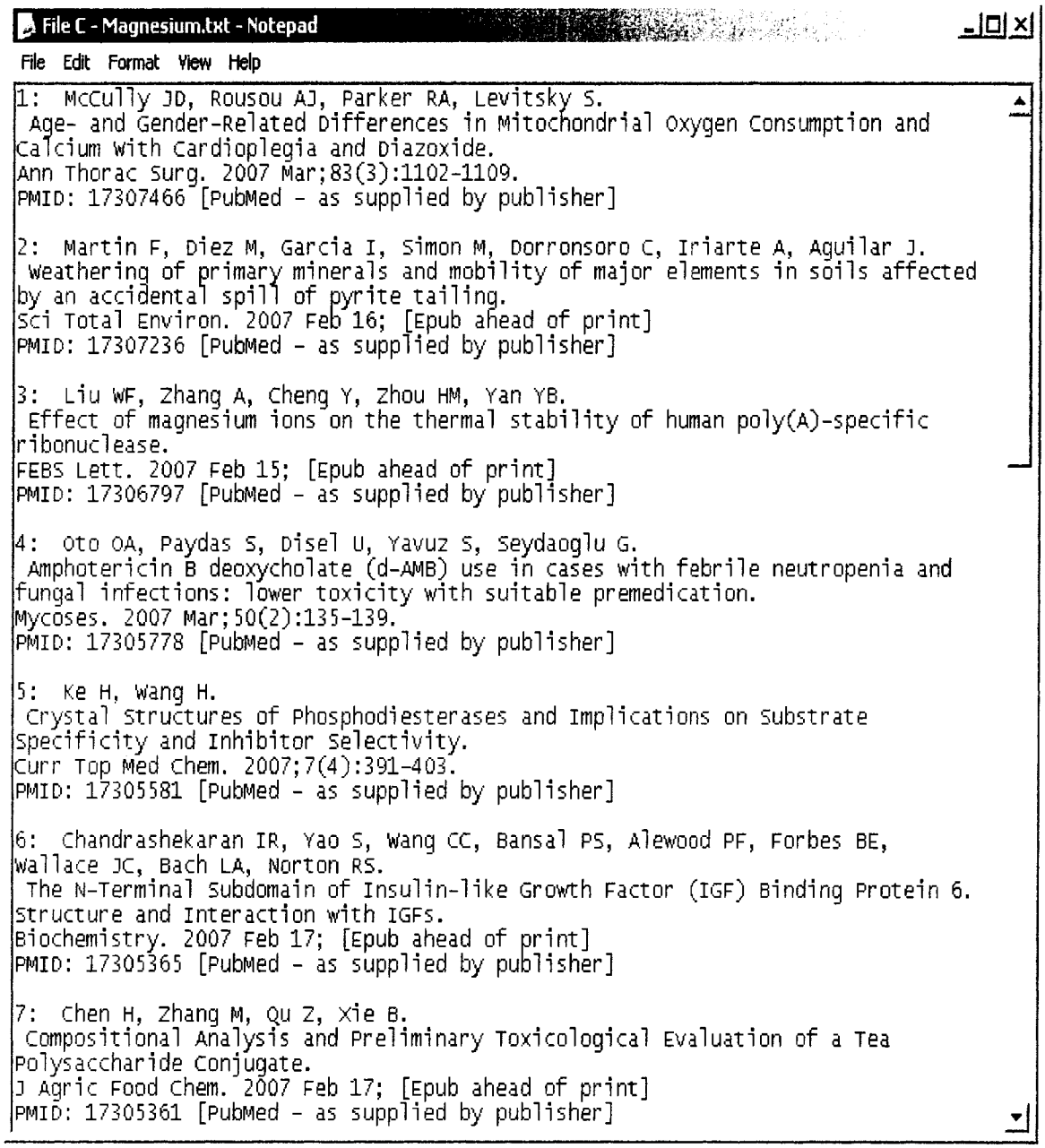

Figure 5.3: Sample display of File C articles of title Magnesium

These File A and File C were uploaded to the LBD tool and further processing of stopword lists and stemming are applied. The next figure shows the corresponding B-list produced by the program.

Rank	Prob	B-term
1	0.93	rofecoxib
2	0.93	lamotrigine
3	0.93	ketorolac
4	0.93	gabapentin
5	0.93	transcranial doppler
6	0.94	metoprolol
7	0.93	aneurysmal subarachnoid
8	0.93	mitral valve prolapse
9	0.93	topiramate
10	0.93	risk stroke
11	0.93	cardiovascular risk factor
12	0.93	orexin
13	0.93	5-ht3 receptor
14	0.93	channel opener
15	0.93	neuropathic pain
16	0.93	atherosclerosis risk community
17	0.93	isradipine
18	0.93	5-htla receptor
19	0.93	nmda receptor antagonist
20	0.93	near infrared spectroscopy

Figure 5.4: B-list of title words and terms common to File A & C.

5.2.2. Relevant AB and BC Literature display

The figure shows the selected entries from a printed display of the output for two sets of titles that contain the Target literature (A-left column) or Source literature (C-right column), respectively. The headings shown are B-terms (in alphabetic sequence). Under each B- term are titles containing that term in the two columns of the display. This arrangement facilitates comparing File A titles with File C titles for each B-term that they have in common. Shown here are examples of B-terms in which biologically meaningful relationships A-B and B-C are suggested by the titles; citations to the literature arc also

identified. Although, for each B-term, only a few titles are shown, in general there may be as many as several hundred. To avoid excessive output, a limit to the number of titles displayed for any one B-term can be set interactively by the user [38].

AB literature	B-term	BC literature
"migraine disorders"[MeSH Term...	cardiovascular risk factor	"magnesium"[MeSH Terms] OR mag...
1: **Cardiovascular risk factors** and migraine: the GEM population-based study. 2005 Add to clipboard		1: [Magnesium, **cardiovascular risk factors** and atherosclerosis] 2005 Add to clipboard
2: **Cardiovascular risk factors** associated with migraine. 2005 Add to clipboard		2: [Effects of two Sicilian red wines on some **cardiovascular risk factors**] 2004 Add to clipboard
3: **Cardiovascular risk factors** and migraine: the GEM population-based study. 2005 Add to clipboard		3: Magnesium deficiency in African-Americans: does it contribute to increased **cardiovascular risk factors**? 2003 Add to clipboard
4: The influence of genetic and **cardiovascular risk factors** on the CADASIL phenotype. 2004 Add to clipboard		4: Nutritional variation and **cardiovascular risk factors** in Tanzania--rural-urban difference. 2003 Add to clipboard
5: Headache and **cardiovascular risk factors**: positive association with hypertension. 1999 Add to clipboard		5: The influence of calcium and magnesium in drinking water and diet on **cardiovascular risk factors** in individuals living in hard and soft water areas with differences in cardiovascular mortality. 2003 Add to clipboard
		6: Dietary minerals and modification of **cardiovascular risk factors**. 2003 Add to clipboard

Figure 5.5: AB-BC literature display for a B-term "Cardiovascular risk factor"

5.3. Weightage Formula for Ranking

The subject-heading weight for a given title B-term is:

Sh-wt = 100*ncom / (nAB*nBC).

This expression represents the density of ncom among all possible pairs of titles displayed (AB with BC), hence, the multiplicative denominator AB*BC [9]. Pairs are the most cogent units to count because the purpose of the display is to facilitate the recognition of potential complementary relationships between A-titles and C-titles. This weight is used to rank title B-terms, placing the higher weights at the top.

For a given B list term [9]

- {AB} = subset of records in A containing that title-term.
- {BC} = subset of records in C containing that title-term.
- nAB = number of records in {AB}
- nBC = number of records in {BC}
- ncom = the number of unique subject headings that {AB} and {BC} have in common.
- Weight for a given title: B-term = 100*ncom / (nAB*nBC).

where,

ncom is the number of unique subject headings that {AB} and {BC} have in common.

nAB is the number of records in the set {AB} and

nBC is the number of records in the set {BC}.

5.4. Complexities with LBD Process

Scientific arguments in general cannot be extracted automatically from titles, abstracts, or the full text of articles, but titles often can serve as pointers or clues that

guide the viewer to arguments presented in the text. For that reason, Arrowsmith provides a link from each B-term to the A and C titles from which it was extracted, and so helps the user assess whether it might qualify as a target [9]. 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 [3].

Discovery of 'undiscovered public knowledge' can result from syntaxes of concepts in any one of six enumerated categories related to the levels of scientific activity:

a. Hidden refutations or qualifications of hypotheses,

b. Inferences from transitive relations,

c. Cumulative weak tests,

d. Unrecognized or hidden analogies,

e. Hidden correlations, and

f. In general, recombination of levels of signification (linking previously published data and information to new knowledge paradigms).

These approaches substantiate the view that objective scientific knowledge constitutes a universe open to exploration and further discovery [1].

5.5. The Role of Human Intelligence

At several points in the procedure, the LBD tool receives a boost from a human input that helps it perform as if it were intelligent. The first boost is the choice of the

problem and its literature C, plus the choice of A as a specific target. Using A and C to conduct a good search also require knowledge, experience, and judgment at the outset. The second boost is the stoplist filter, which greatly reduces the number of useless connections that otherwise would clutter the output.

The stoplist is compiled using human judgment (and guesswork) concerning which words probably could not play any useful role in forming biologically meaningful and helpful linkages. The remaining boosts come from the user in editing the B-list and the A-list and in forming groups. Finally, given the juxtaposed AB-BC titled or abstracts, any identification of promising implicit linkages of scientific importance depends on the knowledge and perspicacity of the user [15].

5.6. Information Retrieval and Text Mining Methods to Aid LBD tools

Most methods and developments can be divided into five distinct text mining steps. The steps are:

- Text gathering,
- Text preprocessing,
- Data analysis,
- Visualization,
- Evaluation.

The goal of this thesis was therefore to analyze the different methods applicable to the five steps of text mining process. The emphasis lies on the computer science focused solutions, as this is the area of our expertise. Beside the task of integrating existing methods into the framework and providing interfaces for them, the individual research tasks for each step split up. In text gathering our research focus lies on PDF

conversion, as it seems to be a bottleneck in that area and it is also frequently needed for text mining in full text [30].

CHAPTER 6

CONCLUSION AND FUTURE WORK

Literature Based Discovery (LBD) has increasingly become the focus of research in knowledge discovery in textual data [3]. For a wider deployment of LBD tools, more research and development are needed to optimally display the discovery results. The results are, possibly very diverse, pieces of knowledge that have to be combined and integrated by the users themselves [8]. There is also a critical need to develop theoretical foundations for LBD and also fully automate the LBD process by employing "Supervised machine learning algorithms" [26].

A decade ago, the thought that information retrieval would be such a prominent technology occurred to very few individuals. Today, some of their ideas have found every day use in the search engines that are vital for successful negotiation of the Web. Perhaps a decade from now, discovery-support tools will be equally prevalent [10].

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 [3].

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 [3].

BIBLIOGRAPHY

1. F E Morrisay. "Theory and practice of knowledge discovery in scientific, technical and medical databases", Reference http://conferences.alia.org.au/shllc2001/papers/morrissey.html
2. Swanson DR. "Undiscovered public knowledge". *Library Quarterly* 1986; 56: 103-118.
3. "Recent Advances in Literature Based Discovery", Reference http://www.dimacs.rutgers.edu/~billp/pubs/JASISTLBD.pdf
4. "Undiscovered Public Knowledge", Reference http://www.iawiki.net/UndiscoveredPublicKnlowlegde
5. Larry S. Jackson (23 January 2002) "Supercomputing Detection of Swanson's Relationship between Raynaud's Disease and Dietary Fish Oil" reference http://realfun.isrl.uiuc.edu/sc-upk/papers/UIUCLIS_2002_2_UPK.html
6. Swanson, D. R. (1986), "Fish oil, Raynaud's syndrome, and undiscovered public Knowledge", *Perspect. Biol. Med.*, Vol. 30(1), pp. 7–18.
7. Hearst MA. "Untangling text data mining". Proc. ACL 1999.
8. Weeber M, Kors JA, Mons B. "Online tools to support literature-based discovery in the life sciences. Brief Bioinform". 2005 Sep; 6(3):277-86.
9. Swanson DR, Smalheiser NR, Torvik VI. "Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH)". *JASIST* 2006; 57(11):1427-1439.

10. Michael Gordon, Robert K.Lindsay, Weiguo Fan "Literature Based Discovery on the World Wide Web". *ACM Transactions on Internet Technology*, Vol. 2, No. 4, November 2002, Pages 261–275.

11. Weeber M, Vos R, Baayen RH. "Using concepts in literature-based discovery: Simulating Swanson's raynaud - fish oil and migraine - magnesium discoveries". *JASIST* 2001; 52: 548-557.

12. Dimitar Hristovski, Borut Peterlin, Sašo Džeroski, Janez Stare. "Literature Based Discovery Support System and its Application to Disease Gene Identification".

13. Koike A, Takagi T, "Knowledge discovery based on an implicit and explicit conceptual network". *J.Am. Soc. Inf. Sci. Tech., in press.* 58(1): 51-65, 2007.

14. Medline. Reference. http://www.nlm.nih.gov/pubs/factsheets/medline.html

15. Swanson DR, Smalheiser NR. "Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery". *Library Trends* 1999; 48: 48-59.

16. Meredith M Skeels, Kiera Henning, Meliha Yetisgen Yildiz, and Wanda Pratt. "Interaction Design for Literature-based Discovery". CHI 2005, April 2–7, 2005, Portland, Oregon, USA. *ACM 1-59593-002-7/05/0004.*

17. Raul E. Valdes-Perez. "Principles of human-computer collaboration for knowledge discovery in science. Artificial Intelligence". *Artificial Intelligence* 107 (1999).335-346.

18. Raul E. Valdes-Perez "Discovery Tools for Science Apps". *Communications of the ACM*, November 1999/Vol. 42, No. 11.

19. Smalheiser NR, Swanson DR. "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses". *Computer Methods and Programs in Biomedicine* 1998; 57: 149-153.

20. Smalheiser NR, Torvik VI, Bischoff-Grethe A, Burhans LB, Gabriel M, Homayouni R, Kashef A, Martone ME, Perkins GA, Price DL, Talk AC, West R. "Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators". *J Biomed Discov Collab.* 2006 Jul 3; 1(1):8.

21. Roger Hale. "Text Mining: Getting more value from literature resources". *Published in Drug Discovery Today,* Volume 10, Issue 6, Pages 377-379 (15 March 2005).

22. Using Semantic Content for Formulating and Assessing Hypotheses. Reference. http://www.mindswap.org/webai/2002/fall/Concept_20Bridging_20on_20the_20Semantic_20Web.html

23. Aaron M. Cohen and William R. Hersh. "A survey of current in biomedical text mining". *Henry Stewart Publications 1467-5463. Briefings in Bioinfomatics.* Vol 6. No 1. 57-71. March 2005.

24. Van der Eijk, C. C., van Mulligen, E. M.,Kors, J. A. et al. (2004), "Constructing an associative concept space for literature-based discovery", J. *Amer. Soc. Inf. Sci. Tech.,* Vol. 55(5), pp. 436–444.

25. Dimitar Hristovski1, Borut Peterlin2 "Improving Literature Based Discovery Support by Background Knowledge Integration".

26. Robert Finn. "Program Uncovers Hidden Connections In The Literature". Reference http://www.the-scientist.com/article/display/18032/.

27. Smalheiser NR, Swanson DR. "Linking estrogen to Alzheimer's Disease: an informatics approach. Neurology". 1996; 47: 809-810.

28. C. Chibelushi, B. Sharp, A. Salter (2004). "A Text Mining Approach to Tracking Elements of Decision Making: a pilot study". *In Sharp, B. (ed.) Proceedings of 1st International Workshop on Natural Language Understanding and Cognitive Science (NLUCS2004) collocated with ICEIS 2004. Porto, Portugal*, 13 April 2004, pp. 51-63. ISBN: 972-8865-05-8.

29. A Literature-based Approach to Scientific Discovery. Reference http://arrowsmith2.psych.uic.edu/cci/workshop/swanson.ppt

30. Wei Huang, Yoshiteru Nakamoria, Shouyang Wang and Tieju Ma. "Mining Scientific Literature to predict new relationships". *Intelligent Data Analysis* 9 (2005) 219–234 219. IOS Press.

31. Catherine Blake, Carryon Anderson "The shift from information retrieval to synthesis Theme: The grand challenges".

32. Johannes Stegmann and Guenter Grohmann. "Transitive text mining for information extraction and hypothesis generation".

33. ARROWSMITH linking documents, disciplines, investigators, and databases. Refernce. http://kiwi.uchicago.edu/webwork/PURPOSE.html.

34. Porter Stemming Algorithms Reference, http://www.tartarus.org/~martin/PorterStemmer/index.html.

35. An algorithm for suffix tripping. M.F.Porter. 1980. Reference http://www.tartarus.org/~martin/PorterStemmer/def.txt.

36. Srinivasan P. "Text mining: generating hypotheses from Medline". *Journal of the American Society for Information Science and Technology.* 2004; 55:396–413.

37. Swanson DR. "Medical literature as a potential source of new knowledge". *Bull. Med. Libr. Assoc.* 1990; 78: 29-37.

38. Swanson DR, Smalheiser NR. "An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence". 1997; 91:183–203.

39. Langley, P. "The computer-aided discovery of scientific knowledge". *In Proceedings of the 1st International Conference on Discovery Science* (1998). Springer New York.

40. Pratt, W., and Yetisgen-Yildiz, M. "LitLinker: Capturing Connections across the Biomedical Literature". In Proc. K-Cap'03, Florida (2003), 105-112.

41. Swanson DR. "Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy". *J Am Soc Inf Sci.* 1989 Sep; 40(5):356–358.

42. Swanson, DR. "Complementary structures in disjoint science literatures". *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Chicago, ACM Press; 1991. pp. 280–289.

43. Brigitte Mathiak and Silke Eckstein."Five Steps Text Mining in Biomedical Literature". *In Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, held in Conjunction with ECML/PKDD* in Pisa, Italy. 24 September 2004.

[44] Smalheiser, N.R. (2001) "Predicting emerging technologies with the aid of text-based data mining: a micro approach". *Technonvation* 21: 689-693.

VITA

Graduate College
University of Nevada, Las Vegas

Ramalakshmi Sundar

Home Address:
4224 Cottage Circle,
4
Las Vegas, Nevada 89119

Degrees:

Bachelor of Engineering, Electronics & Communication Engineering, 2003
University of Madras

Master of Science, Computer Science, 2007
University of Nevada, Las Vegas

Special Awards and Honors:

- "Best outgoing student" – Awarded in 2004 by Department of Electronics & Communication Engineering, India.
- Certificate of Merit & Cash Award from the Collegiate Education Department, India for Commendable Performance in the High School Board Examination.

Thesis Title: Literature Based Discovery: Techniques and Tools

Thesis Examination Committee:

Chairperson, Dr. Kazem Taghva, Ph. D.
Committee Member, Dr.Thomas Nartker, Ph. D.
Committee Member, Dr. Ajoy Datta, Ph. D.
Graduate Faculty Representative, Dr. Shahram Latifi, Ph. D.