Literature-based knowledge discovery using an associative concept space

B.F. van der Lans

March 18, 2004

Contents

1 Introduction **1**
- 1.1 Scientist's fields of interest . . . 1
- 1.2 Undiscovered public knowledge . . . 2
- 1.3 Searching for undiscovered public knowledge . . . 4
- 1.4 Goal, methodology, and structure of this thesis . . . 5

2 Resources and tools for literature-based knowledge discovery **7**
- 2.1 Availability of scientific literature . . . 7
- 2.2 Extracting knowledge from text . . . 8
- 2.3 Resources and tools in the biomedical domain . . . 9
  - 2.3.1 Electronic availability of biomedical literature . . . 10
  - 2.3.2 Resources used to interpret biomedical literature . . . 11
  - 2.3.3 Making concept representations of text . . . 13

3 Literature-based knowledge discovery systems **17**
- 3.1 A general architecture . . . 17
- 3.2 Literature-based knowledge discovery systems . . . 20
  - 3.2.1 Swanson and Smalheiser . . . 20
  - 3.2.2 Gordon and Lindsay . . . 22
  - 3.2.3 Weeber et al . . . 23
  - 3.2.4 Hristovski et al . . . 24
  - 3.2.5 Srinivasan . . . 25
- 3.3 General issues in literature-based knowledge discovery . . . 27
  - 3.3.1 Knowledge source . . . 28
  - 3.3.2 Knowledge representation . . . 29
  - 3.3.3 Open and/or closed discovery process . . . 30
  - 3.3.4 Evaluating systems . . . 30

4 System description **32**
4.1 Associative Concept Space 32
4.2 Two-step approach 35
4.3 One-step approach 36
4.4 Modifications 37

5 Method of evaluation **42**
5.1 Test case 42
5.1.1 Description 43
5.1.2 Usage in the evaluation 44
5.2 ROC analysis 45
5.3 Evaluation methodology 50
5.3.1 Two-step approach 50
5.3.2 One-step approach 52

6 Test results **54**
6.1 Two-step approach 54
6.1.1 Experiment 1-1: The basic approach 55
6.1.2 Experiment 1-2: Combining rankings 56
6.1.3 Experiment 1-3: Using a more general document set . 58
6.1.4 Experiment 1-4: Using inverse document frequency . . 61
6.1.5 Experiment 1-5: Using semantic categories 61
6.2 One-step approach 62
6.2.1 Experiment 2-1: The basic approach 62
6.2.2 Experiment 2-2: Using inverse document frequency . . 64
6.2.3 Experiment 2-3: Using semantic categories 65
6.3 Summary of results 66

7 Discussion and outlook **68**
7.1 Discussion of results 68
7.2 Limitations of our evaluation method 71
7.3 Suggestions for further research 72
7.4 Outlook 73
7.5 Acknowledgements 74

Bibliography **76**

Appendix **81**

Chapter 1

Introduction

1.1 Scientist's fields of interest

Science can be described as "a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through scientific method." The goal of scientists is to contribute new knowledge to science. Scientists do this by the use of scientific method, which can be described as "principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses" (Merrian-Webster, 2004).

In order to be able to formulate problems and hypotheses, scientists study the existing pool of scientific knowledge. The size of this pool is immense and expands at an exponential rate. An example from the medical domain is the number of medical journals, which has doubled every 19 years since 1870 (Wyatt, 1991). As a consequence, no single scientist is capable of studying all available scientific knowledge. This forces scientists to focus on a part of this knowledge, to specialize.

We will call the part of scientific knowledge a specific scientist focuses on and which he studies and keeps up with his *field of interest*. This field of interest might be very specialized in a certain discipline of science, or it might be more general and span multiple disciplines. Either way, it is limited by the *knowledge capacity* of the individual scientist. A scientist can only absorb a limited amount of knowledge and can only keep up with a limited amount of new knowledge.

New knowledge spreads by communication among scientists. Communication is more frequent between scientists with overlapping fields of interest.

This communication can take many forms. There is direct communication, such as telephone calls, e-mails, and discussions at meetings and conferences. More indirect and more formal communication takes the form of publishing in journals and books.

A scientist tries to study and keep up with the knowledge in his field of interest. To do this, he must determine which instances of communication to use as a source of knowledge for his field. For example, he must choose which books to read and which conferences to attend. This process is not trivial. Different scientists have different fields of interest and a scientist is only interested in those parts of the fields of interest of others that overlap with his own field. Also, fields of interest can change over time. A failure to determine correctly which knowledge sources to study and keep up with means that a scientist will not have all knowledge available in his field of interest.

Some of the knowledge missed may be relevant to a current research topic of the scientist. A scientist without knowledge of all relevant new information on his research topics might make less progress than he would have if he did have this knowledge. Also, different scientists might be solving the same problem without being aware of each other's efforts.

Another way in which relevant knowledge can be missed, is if it is outside a scientist's field of interest. This should not be possible, as an ideal field of interest would include all knowledge relevant to the scientist's current research. In practice, a scientist's field of interest is that part of scientific knowledge he studies and keeps up with. A scientist might not be aware of existing knowledge of interest to him and thus not include it in his field of interest.

To sum up, scientific progress would benefit if all scientists had all knowledge relevant to their research topics. Part of the reason they do not have all this knowledge are the difficulties involved with determining which knowledge sources to use to keep up with knowledge in their field of interest. Another reason is that some relevant knowledge can only be found outside of their field of interest.

1.2 Undiscovered public knowledge

Science is divided into manageable units, or specialties, and so scientific knowledge is created and assimilated in manageable units. While many of these units of knowledge are related to each other, they are created to some degree independently of each other and the relationships among them may

be unknown to even their creators. Because of the independence with which they are created, the relationships among the units may remain undiscovered. (Swanson, 1986)

This is similar to what we described in section 1.1. A scientist's field of interest includes the units of knowledge he is working on and those units of which he knows that they have a relationship with his own units. His field of interest should include all units related to his own, but this will not be the case if the scientist has no knowledge of some of those related units. The related units not included in his field of interest might contain knowledge relevant to his research.

It might even be so that some progress can only be made when two related different pieces of knowledge are brought together. However, when no single scientist's field of interest includes those two pieces of knowledge, their relationship will remain unknown. Even when a specific scientist's field of interest includes both pieces, the relationship may remain unknown when a scientist chooses the wrong sources of knowledge for that field of interest. Swanson called these unknown relationships "undiscovered public knowledge". *Public* because all pieces of knowledge needed already exist and are publicly available, *undiscovered* because no scientist has brought the pieces together yet.

To illustrate this, let us consider the following situation. A relationship exists between two topics, say A and C. However, no unit of scientific knowledge describes this relationship, and thus it is unknown. However, there might be a unit describing a relationship between A and another topic B. Also, there exists another unit, in which a relationship between B and C is described. A scientist with a field of interest including both the unit describing the relationship between A and B and the unit describing the relationship between B and C might notice both relationships. He might then hypothesize that there is a relationship between A and C. If he succeeds in supporting this hypothesis with (experimental) evidence, new knowledge is discovered. Figure 1.1 shows an example where a scientist's field of knowledge does not include both units and the relationship will remain undiscovered.

Don R. Swanson has made several discoveries by studying separate units of knowledge. For example, he discovered a relationship between two topics, Raynaud's disease and fish oil, while studying two separate units of knowledge. In one of them, fish oil A has been proven to improve blood circulation B. Patients with Raynaud's disease C have intermittent blood flow B in their extremities. Both of these relationships are supported by substantial scientific evidence and literature. Taken together, these relationships sug-

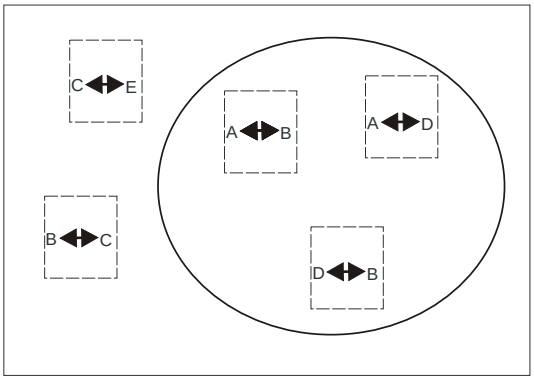

Figure 1.1: Example of an undiscovered relationship between some A and some C. The circle represents the field of interest of some scientist interested in A. The squares are units of knowledge. The scientist will not discover the relationship of A with C through B, because the unit describing a relationship between B and C is outside of his field of interest.

gest that dietary fish oil has a relationship with Raynaud's disease, namely that a diet containing fish oil can ameliorate the effects of Raynaud's disease. (Swanson, 1986). This relationship was not known at the time of Swanson's discovery, but has later been shown to exist in a clinical trial (Chang et al., 1988; DiGiacomo et al., 1989). The hypothesis could be proven and therefore new knowledge was discovered.

1.3 Searching for undiscovered public knowledge

When we wish to improve scientific progress by finding undiscovered public knowledge, we will need to search in the pool of existing public knowledge. More specifically, we will need to bring together two parts of existing knowledge that are relevant to the same research topic and have not been brought together before. An important decision in this process is choosing what exactly will serve as a source of knowledge.

As stated before, there are many forms of communication among scientists; all representing exchanges of knowledge. The recorded forms of communication can serve as a source of knowledge when operationalising a discovery process to unearth undiscovered public knowledge. There are many of these recorded forms, but we will focus on the publications made

by scientists in journals and books. We will refer to these publications as *the scientific literature*.

The scientific literature has several advantages over other sources of knowledge. Firstly, it has been reviewed by peer scientists and is therefore likely to be more reliable than other sources, such as websites. Secondly, literature covers most existing scientific knowledge. Among scientists, it is the standard way of adding new knowledge to the existing pool. Thirdly, scientific literature is accessible. It is available in libraries, and also increasingly electronically. A disadvantage of scientific literature is the delay with which it is published. Sources such as websites are usually far more up-to-date. Considering both the advantages and disadvantages, the scientific literature seems a reasonable representation of current existing scientific knowledge.

When we wish to search for undiscovered public knowledge, we will search for it in the scientific literature. A scientist reads the literature in his field of interest. In this way, chance discoveries such as Swanson's discovery from section 1.2 can be made. A more structured approach to search undiscovered hidden knowledge will require studying much more literature than just the literature in one field of interest. However, fields of interest exist because scientists have limited knowledge capacity. Therefore, scientists will need help to make searching for hidden knowledge in a structured way possible.

Fortunately, several developments might assist in providing this help. One is the already mentioned increased electronic availability of the scientific literature. Others are developments in research in making literature available in a structured way and developments in scientific research in processing, searching and extracting information from text using computers. Combining all these developments, it is possible to use a computer to assist scientists in searching for undiscovered public knowledge. We will call this search process *literature-based knowledge discovery*.

1.4 Goal, methodology, and structure of this thesis

Goal

The main goal of this thesis is to make recommendations for a system that assists scientists in finding undiscovered public knowledge in the scientific literature. This system is a computer system which interacts with the user to exploit the user's knowledge without requiring more than a reasonable

amount of effort from the user.

Methodology

Since Swanson first published the idea of undiscovered public knowledge, several scientists have done research in literature-based knowledge discovery. This research has resulted in several systems for literature-based knowledge discovery. We will begin our study by studying some of the available resources and tools which were used by these systems and/or can be used by our own system. Because there are some excellent resources and tools in the biomedical domain, we will pay special attention to this domain. We will then examine and compare the existing systems for literature-based knowledge discovery in the biomedical domain. From this comparison, we will derive guidelines for our own system.

Following these guidelines, we will suggest two basic approaches to searching for undiscovered public knowledge. We will also suggest possible improvements to these approaches. Our approaches will center around a tool called the Associative Concept Space (ACS), which is described in section 4.1. We will implement our approaches and evaluate them using a case study of the first discovery made by Swanson. During this process, the added value of the ACS in our system will also be evaluated. Finally, we will discuss our findings and do suggestions for further research.

Structure

The structure of the remaining part of this thesis is as follows: chapter 2 deals with tools and resources useful in literature-based knowledge discovery. In chapter 3 previous systems for literature-based knowledge discovery are discussed. Based on the advantages and disadvantages of each of these systems, we will make suggestions for a new system in chapter 4. This chapter will also discuss the tool used in those suggestions. Chapter 5 will then describe the case and evaluation method used to evaluate our suggestions. The results of this evaluation are discussed in chapter 6. Finally, chapter 7 concludes with a general discussion and suggestions for further research.

Chapter 2

Resources and tools for literature-based knowledge discovery

In this chapter, we discuss tools and resources that can be used for literature-based knowledge discovery. In section 2.1, we discuss how the scientific literature is available electronically. Next, we discuss how computers can infer 'knowledge' from text. Section 2.3 discusses resources and tools for literature-based knowledge discovery in the biomedical domain.

2.1 Availability of scientific literature

Our goal is to make recommendations for a system that assists scientists in finding undiscovered public knowledge in the scientific literature. The reason scientists need assistance for this is largely the same as the reason for the existence of undiscovered public knowledge; the amount of literature scientists can handle is only a fragment of the total amount of scientific literature available. To do more than make chance discoveries, much more literature than one scientist can handle needs to be searched. Since computers can handle large amounts of data easily, a literature-based knowledge discovery system should at least be partly automated.

To examine how computers can help with literature-based knowledge discovery, we need to consider how the scientific literature is available to automated systems. The scientific literature consists of the publications made by scientists. These publications are published in journals and books. However, to be of use to an automated system, they should be available

electronically.

Luckily, more and more literature is available online. Some scientists have begun to publish their articles in online journals. Also, many traditional journals offer online access to their content. An example is LinkOut, an extensive list of biomedical journals who provide online access to their content (National Library of Medicine, 2003a). Note, however, that online access to the full text of articles is still limited and when available, often not free of charge.

Access to the titles and abstracts of scientific publications is more readily available. Most libraries today offer free online searching of their catalogue. Also, many libraries use some form of indexing. With indexing, keywords are assigned to articles describing their contents. Using the online search functions of libraries results in online access to titles, abstracts and keywords describing the contents of articles. These sources can be easily accessed by automated systems.

2.2 Extracting knowledge from text

Access to the scientific literature is not all we need for literature-based knowledge discovery. To a computer, the text of which the titles, abstracts, and full text of articles consists is just a string of characters. To a human, that same piece of text has meaning. To humans, the *terms* in the text - consisting of one or more words - refer to concepts. A concept is "something conceived in the mind, a thought, a notion" (Merrian-Webster, 2004).

The relationships existing between terms and concepts are complex. For one thing, concepts exist in the mind and differ from person to person. Another problem is that the relationships are not one-on-one. Different terms, *synonyms*, may refer to the same concept. Also, the same term, a *homonym*, may be used to describe different concepts. Figure 2.1 shows examples of a homonym and a synonym.

If a computer would be able to recognise concepts in text, it would be able to handle it in a more meaningful way than just as a string of characters. Because of the complex relationships described above, the process described above is not trivial.

Thesauri have been developed to deal with these complexities. A thesaurus is a structured list of concepts. These concepts have unique numbers associated with them and all the terms that are used to refer to them. These terms include synonyms, lexical variants and translations. A thesaurus also contains hierarchical relations between the concepts contained in it. These

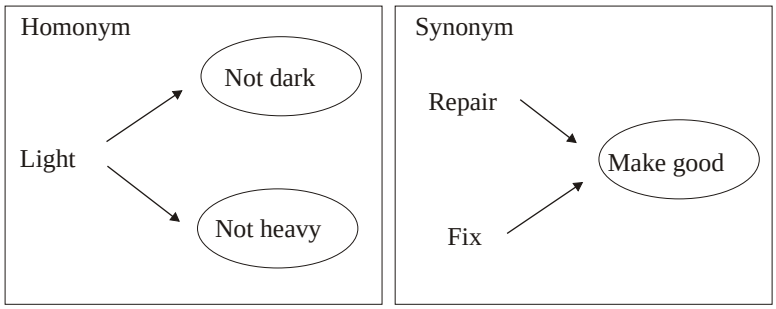

Figure 2.1: Examples of a homonym and a synonym. The circled words represent concepts, the others represent terms.

are of the is-a form, where a more specific concept is an instance of a more general concept. (National Library of Medicine, 2003b)

The concepts in the thesaurus are defined clearly, making the relationships between terms and concepts less person dependant. Also, synonyms can be easily matched with only one concept. The problem of homonymity, one term used for more than one concept, is not solved by thesauri. Contextual information is needed to determine which concept is referred to.

A resource for the non-hierarchical relationships among concepts is a *semantic network*. A semantic network records concepts and the relationships among them. Another resource that can be used to interpret text is a *lexicon*. A lexicon contains the terms in a language and their definitions. (National Library of Medicine, 2003b)

2.3 Resources and tools in the biomedical domain

Several practical resources and tools for literature-based knowledge discovery in the biomedical domain are listed here. The first deals with the electronic availability of medical literature and the structured ways in which this is available. Next, some resources that can be used to interpret biomedical text are discussed. We finish with a discussion of two tools that make concept representations of text.

Figure 2.2: An example of a MEDLINE record. (National Library of Medicine, 2003b).

2.3.1 Electronic availability of biomedical literature

A major contributor to the electronic availability of the medical literature is the United States National Library of Medicine (NLM). The NLM is the world's largest medical library and has a multitude of projects to make information easily available by electronic means. The most important of these projects is MEDLINE.

"MEDLINE is the premier bibliographic database of the NLM. It contains over 12 million references to articles published between 1966 and now. It covers basic biomedical research, clinical sciences, and life sciences that are critical to biomedical research." (National Library of Medicine, 2003b).

A typical MEDLINE record includes the title, author, and publishing information of an article. It may also contain an abstract of the article. Figure 2.2 shows an example. The records contained in MEDLINE can be accessed online by PubMed, free of charge. Through PubMed, MEDLINE can be searched using search terms such as author names, title words, text words or phrases, journal names, or combinations of these.

Many of the articles contained in MEDLINE are indexed by human in-

dexers. They assign terms to the records which describe what each document is about. These descriptors are chosen from a structured list called Medical Subject Headings (MeSH). The descriptors in MeSH are structured in a hierarchy. They also have cross-references among them. MeSH contains almost 22,000 descriptors. The MeSH descriptors assigned to a record are divided in major and minor MeSH descriptors. The major descriptors describe the most important topics in an article, the minor the other topics.

MEDLINE is essential to literature-based knowledge discovery in the biomedical field. It is used by all of the knowledge discovery systems discussed in the next chapter.

2.3.2 Resources used to interpret biomedical literature

The NLM's Unified Medical Language System (UMLS) project provides a large thesaurus, the UMLS Metathesaurus, and two related resources. These are the SPECIALIST Lexicon and the UMLS Semantic Network.

UMLS Metathesaurus

The UMLS Metathesaurus is compiled from many already existing thesauri in the biomedical field. It preserves the information found in those thesauri, adds certain basic information and establishes new relationships among terms found in different thesauri. The July 2003AB edition includes 900,551 concepts and 2.5 million concept names derived from over 100 biomedical source thesauri. A typical record in the UMLS Metathesaurus consists of a name, a number, a definition, synonyms, and translations. It also contains references to the source thesauri and to ancestors in the hierarchy. Figure 2.3 shows an example of a Metathesaurus entry.

SPECIALIST lexicon

The SPECIALIST lexicon is intended to be a general English lexicon that is augmented with many biomedical terms. Lexical entries may be single- or multi-word terms. The information associated with an entry includes syntactic category (i.e., verb, noun, etc.), inflectional variation (e.g., singular and plural for nouns, the conjugations of verbs), and allowable complementation patterns (i.e., the objects and other arguments that verbs, nouns, and adjectives can take).

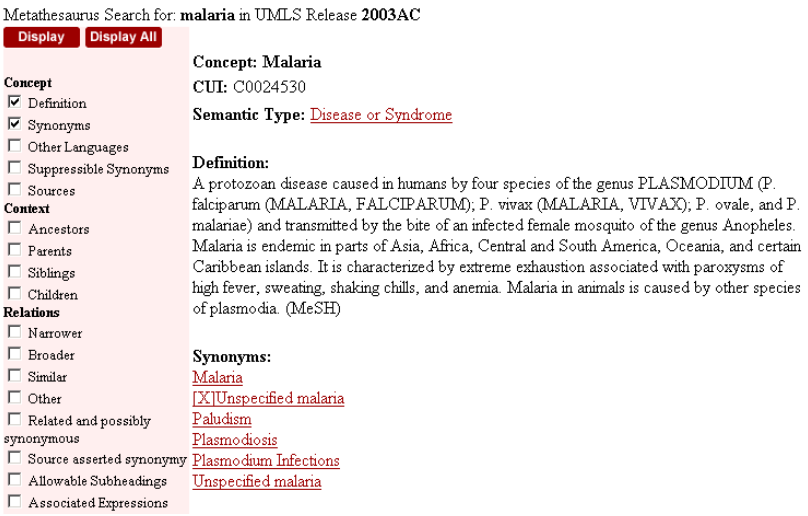

Figure 2.3: An example of a Metathesaurus entry. Only part of the entry is displayed here. (National Library of Medicine, 2003b)

UMLS semantic network

The UMLS semantic network categorizes all the concepts in the UMLS Metathesaurus. Concepts in the Metathesaurus have one or more of the 134 semantic types in the network assigned to them. There are 54 links between these semantic types, which represent important relationships in the biomedical domain. The semantic network also contains information regarding each of the semantic types.

2.3.3 Making concept representations of text

This section describes two tools that have been developed to identify concepts in natural language text: MetaMap and Collexis. We first discuss MetaMap, then Collecis, and we conclude with a brief comparison.

MetaMap

MetaMap is a program developed at the NLM to map biomedical text to concepts in the UMLS Metathesaurus (Aronson, 2001). There is a Java implementation of MetaMap, MMTx, that can use any thesaurus for this, instead of only the UMLS Metathesaurus. The algorithm consists of five steps.

The first step is *parsing*. In this step, a piece of free text is parsed into a list of simple noun phrases. This step uses the SPECIALIST lexicon to recognise phrases in the text. The parser also indicates which part of each phrase is the most central part, the *head*.

The next step is *variant generation*. For each phrase in the list, *variants* are generated. A variant is a phrase word, along with all its acronyms, abbreviations, synonyms, derivational variants, meaningful combinations of these, and inflectional and spelling variants. This step uses the SPECIALIST lexicon and an additional database of synonyms.

After this, a candidate set of all the terms used in the UMLS Metathesaurus to refer to concepts is retrieved, where each retrieved string contains at least one of the generated variants. This step is called *candidate retrieval*.

Next is *candidate evaluation*. In this step candidates are evaluated against the input text. First, a mapping is computed from the phrase words to the candidate's words. Then, the strength of the mapping is computed using an evaluation function consisting of a weighted average of four metrics: *centrality*, *variation*, *coverage*, and *cohesiveness*. These metrics measure respectively the involvement of the head, the variation between phrase and

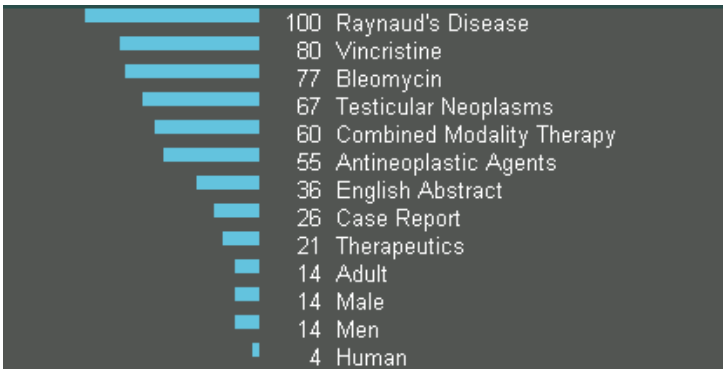

Figure 2.4: Example of a fingerprint in which Raynaud's Disease is the most important concept.

candidate, how much of a candidate matches the text, and in how many pieces. The candidates are ordered according to strength.

The last step is *mapping construction*. Complete mappings are made by combining candidates that were involved in disjoint parts of a phrase. The strength of these combinations is calculated in the same way as the strength of the candidates was. The highest scoring complete mapping is the concept representation of the phrase.

After these five steps, MetaMap has mapped a piece of free text to a list of concept representations of each phrase in the text.

Collexis

In (van Mulligen et al., 2000), Erik van Mulligen and his colleagues describe a system that derives a weighted profile of a scientist from a set of documents by that scientist. Part of this system is an algorithm that maps a given piece of free text to a *fingerprint*. A fingerprint is a weighted list of concepts. The weight of each concept in the fingerprint represents two things. First, a concept with a higher weight is more probable to be the actual concept that the words in the text refer to. Second, a concept with a higher weight has more importance in the text. Figure 2.4 shows an example of a Collexis fingerprint.

The algorithm starts by normalizing the words in the text. The terms in the thesaurus are also normalized. The normalized words from the piece of text are then matched against the normalized words from the thesaurus.

The concepts in the thesaurus whose terms match against words in the text form a list of candidate concepts. The next step is clustering the words found in the text. Words within a certain distance (in words) of each other that refer to the same concept are combined. For each of the concepts, statistical features are computed and combined to form a weight. Six statistical features are calculated.

The first is *specificity*. For each word that both occurs in the text and is used in one of the terms listed in the thesaurus for a candidate concept, the algorithm calculates to how many concepts in the thesaurus that word also could refer. A candidate concept receives more weight if the text contains words that occur in one of that candidate's terms, but do not often occur in terms of other concepts.

The second feature is *similarity*. It measures the fraction of a candidate concept's name that is covered by words from the text. The more of a concept's name is covered by words in the text, the higher weight it receives.

For the third measure, *co-occurrence*, the algorithm looks at each possible pair of candidate concepts. The measure indicates how often each pair of candidate concepts co-occur in other texts. To determine this, resources such as MEDLINE are used. If two candidates often co-occur in other texts, both receive more weight.

Dispersion is the fourth measure. It is the mean distance (in words) between words referring to one concept. A candidate concepts receives more weight if the words referring to it appear close together.

A word, or cluster of words, usually refers to more than one candidate concept (homonymity). The number of concepts that each word or cluster refer to is called the *cluster size*. Candidate concepts receive more weight if the cluster sizes of the words or clusters referring to them are small.

The last measure, *frequency*, represents the number of words or concepts referring to a candidate concept. Candidates with high frequency receive more weight.

When the six measures for each concept have been combined for a single weight for each concept, the concepts are sorted by weight and together form a fingerprint.

Comparison

There are several differences between these two methods to map text to concepts:

- MetaMap uses syntactical information through its recognition of noun

phrases. Collexis does not use this information. This potentially makes MetaMap more accurate.

- MetaMap generates variants from the words in the text and compares these with the concept records in the thesaurus. Collexis normalizes both the words in the text as the words in the thesaurus. MetaMap potentially recognizes more concepts in the text in this way, but also generates more noise.
- Collexis is much faster than MetaMap.

Chapter 3

Literature-based knowledge discovery systems

Several scientists, the first being Swanson, have done research in literature-based knowledge discovery. Several systems have been developed to search for undiscovered public knowledge in the scientific literature. These systems differ in approach and in results. Since we wish to make suggestions for a new system, it is important to study the systems developed previously. In this chapter these systems are discussed and compared. In section 3.1, we describe a general architecture for literature-based knowledge discovery systems. The next section discusses the systems developed previously. Finally, section 3.3 discusses general issues we encountered while studying these systems.

3.1 A general architecture

The systems we discuss all roughly fit a general architecture. In this architecture, two steps are discerned in the discovery process. Systems might consist of either or both of these steps. We will call the first step the open discovery process and the second the closed discovery process, terms introduced by Weeber (Weeber et al., 2001).

Open discovery process

The open discovery process starts with the user expressing interest in a certain topic of interest, topic A. Literature concerning this topic is sought. The system then uses this literature to compile a list of topics that are re-

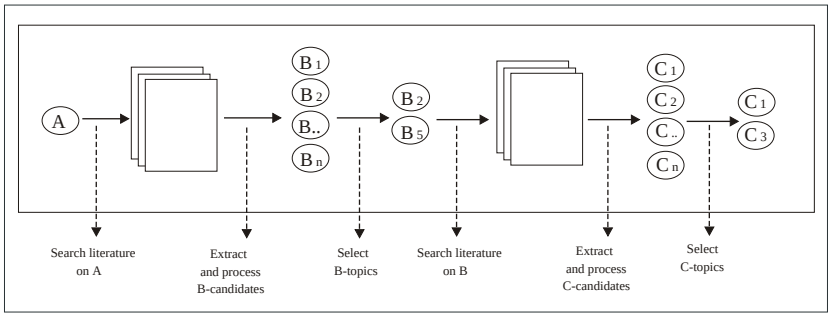

Figure 3.1: Open Discovery Process

lated to topic A. This list generally consists of many topics. Most systems use filtering, ranking, or other processing to put the most interesting related topics forward. The criteria for determining which topics are the most interesting differ strongly among the different systems.

From the resulting list a number of topics is selected, usually by the user. These topics have a relationship with A which is deemed interesting by the user. The selected topics are the B-topics, For these B's, literature is sought. From this literature, a list of candidates for C-topics is extracted. Again, this list is usually very large and again most systems use filtering, ranking, or other processing to put the most interesting related topics forward. On top of that, the list of C-topics is filtered to exclude topic A and topics that have a relationship with A that is already directly described in the literature. The open discovery process is illustrated in figure 3.1.

The C's discovered in this way represent topics that have a potentially relevant relationship with the starting topic A. This relationship is not described directly in scientific literature, but can be inferred from direct relationships between A and one or more intermediate topics B and direct relationships between those B's and the C discovered. If the relationship discovered this way is meaningful and interesting, it may constitute valuable new knowledge.

Closed discovery process

The closed discovery process starts with a hypothesis of the existence of an unknown relationship between some A and some C. If the closed process

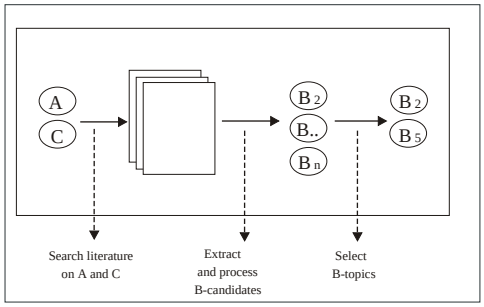

Figure 3.2: Closed Discovery Process

follows an open process, the hypothesis follows from the latter. The closed discovery process can also be used separately from the open process, if a hypothesis already exists.

The system searches for literature that contains either A or C. It then extracts B topics that are directly related with both A and C. Note that the number of B-candidates generated in this way is much smaller than in an open process, were only a relationship with A is required.

The candidates are then filtered and/or ranked to put forward the most interesting ones. Because there are less candidates than in an open process, the criteria for this filtering can be less strict. This means that interesting B-topics can be discovered which were filtered out in the $A - B$ step of the open process. These extra connections between A and C can strengthen the hypothesis resulting from the open discovery process. The closed discovery process is illustrated in figure 3.2.

Summary

To summarize, the open discovery process starts with a topic of interest and finds a topic related to it through an intermediate topic, while the closed discovery process finds (more) intermediate topics connecting two related topics. From the open process follows an initial hypothesis, which can be strengthened by the closed process.

3.2 Literature-based knowledge discovery systems

3.2.1 Swanson and Smalheiser

Background

As mentioned in the introduction, Don R. Swanson made his first discovery in 1986. He accidently discovered a connection between *Raynaud's disease* and *fish oil* after having studied literature on both topics (Swanson, 1986; Swanson, 1987). This discovery led him to identify the existence of undiscovered public knowledge and the opportunity to discover this knowledge by bringing together "complementary but disjoint literatures".

Having identified this opportunity, Swanson began working on a structured approach to searching for undiscovered public knowledge. During this search, he made several other discoveries. In (Swanson, 1988; Swanson, 1989), he describes a connection between *Migraine* and *Magnesium*. A relationship between *Somatomedin C* and *Arginine* is discussed in (Swanson, 1990). Working with Neil R. Smalheiser, he describes relationships between *Magnesium deficiency* and *neurologic disease* in (Smalheiser & Swanson, 1994), between *Indomethacin* and *Alzheimer's disease* in (Smalheiser & Swanson, 1996a), between *estrogen* and *Alzheimer's disease* in (Smalheiser & Swanson, 1996b), and between *Calcium-Independent Phospholipase A2* and *Schizophrenia* in (Smalheiser & Swanson, 1998a).

As Swanson developed his process, he increasingly used informatics tools. Where the first discovery was made completely by hand, the third already made use of several online information sources and tools (Swanson, 1990). This growing use of informatics is also apparent in his proposed systems for literature-based knowledge discovery. His 1991 system for the open discovery process requires a lot of human effort (Swanson, 1991). His 1997 system for the closed discovery process, ARROWSMITH, is more automated. This system is fully described in (Swanson & Smalheiser, 1997; Smalheiser & Swanson, 1998b; Swanson & Smalheiser, 1999).

Swanson evaluated ARROWSMITH by trying to replicate his first three discoveries (Swanson, 1986; Swanson, 1988; Swanson, 1990). He succeeded in replicating his first two discoveries, but failed to replicate the third.

Research

Swanson's research covers both the open and closed discovery process. Topics are represented by words or phrases from the titles of documents. The open process starts with a keyword, representing the starting topic, topic

A. A search is done in MEDLINE to retrieve all article titles containing this keyword. From these titles, all words co-occurring with the keyword are extracted. These represent candidate B-words.

The list of candidate B-words is filtered in four ways. First a pre-compiled stop list is applied. This stop list contains words that are off topic, vague, very general, or otherwise unsuitable as B-words. The second filter only retains words which occur relatively more in titles also containing A than they occur in all the titles in MEDLINE. Thirdly, the user removes all words that are judged to be unsuitable. The fourth filter excludes all words that do not fall into certain user-defined categories of interest.

The resulting list of B-words is used for another search in MEDLINE to obtain all titles containing a B-word. From the result, a list of words that co-occur with a B-word is extracted. These are candidate C-words. First, this list is filtered with the stop list used before. The second filter is also repeated, but based on co-occurrences with the B-words, instead of on co-occurrence with the A-word.

The resulting list of C-words is ranked according to the number of B-words through which they are linked to the A-word. This list is presented to the user, who can then choose C-words for further examination using the closed process. Each of these C-words represents a hypothesis that this C has an relationship with A.

Each hypothesis can be strengthened or rejected by performing a closed discovery process. Before starting the closed discovery process with a certain 'A relates to C' hypothesis, a search in MEDLINE is done to make certain A has no known relationship with C. If a search for MEDLINE records (not just titles) containing both A and C has results, these are studied before starting the closed discovery process.

The closed discovery process, called ARROWSMITH, starts with a given A and C. It retrieves all literature containing either A or C. From this, a list of B-terms is extracted. These B-terms are words or phrases (of up to six words) which co-occur at least twice with both A and C. The stop list is applied next and the list is further edited by the user to remove redundancies and useless terms.

Finally, the system displays each A-B-C link with the titles in which the A-B and B-C co-occurrences are present. The user can now study the context of the co-occurrences and determine which A-C links merit further investigation.

3.2.2 Gordon and Lindsay

Background

Swanson's first discovery was re-examined by Michael D. Gordon and Robert K. Lindsay (Gordon & Lindsay, 1996). They stated that Swanson's work called for a re-examination, due to both its originality and its exploratory, nonsystematic character. The goals they had in re-examining Swanson's work were threefold: to examine and try to replicate the discovery path that led from Raynaud's disease to fish oil; to contribute to research into computer-based tools for supporting literature-based knowledge discovery; and to extend Swanson's findings.

To simulate Swanson's discovery, Gordon and Lindsay used statistical methods related to those in information retrieval. Statistical features extracted from the documents on Raynaud's disease assisted in constructing a query for related documents. With the method developed in this way, they managed to replicate Swanson's discovery.

In (Lindsay & Gordon, 1999), Gordon and Lindsay applied their method to Swanson's second discovery of a connection between migraine and magnesium. The purpose of this experiment was further testing of the applicability of their information retrieval methods to literature-based knowledge discovery. Using the same statistics and method as in their first article, replicating Swanson's second discovery initially failed. Only after applying major changes to the method used, they were able to replicate the discovery.

Research

Gordon and Lindsay used an open discovery process to simulate Swanson's discoveries. Topics are represented by one-, two-, or three-word terms. The process starts with downloading of all MEDLINE records on topic A by searching for term A. The user selects which MEDLINE record fields should be used. From the selected fields, all terms co-occurring with topic A are extracted.

The list can be filtered in up to four ways. First, one or more stop lists can be used. Gordon and Lindsay used three of these lists: one containing frequently used words in English, one with words with high MEDLINE frequency, and one with an ad hoc collection of stop words. The user can select one or more of these and can also add his own. The term list is further shortened by automatically collapsing singular and plural forms of the same term. A third, optional filter is a frequency threshold. This filter removes all words that occur too infrequent, or in too few documents. The

last possibility is the manual collapsing of different terms into one. This is appropriate for synonyms, antonyms, generalizations, and specializations.

The items can then be ranked in four ways, by four different statistics. The first is the *term frequency* $f_{t,X,R}$, which counts how often term X appears in the retrieved record set R. The second statistic is the *document frequency* $f_{d,X,R}$ of X in R. This statistic counts the number of documents in R in which X appears. The statistic *tf·idf* (term frequency · inverse document frequency) is the third:

$$\text{tf·idf} = f_{t,X,R} \cdot \log \frac{N}{f_{d,X,M}}$$

where N is the number of records in the complete record set of MEDLINE M and $f_{d,X,M}$ is the document frequency of X in M. The last statistic used is the *relative frequency* of X in A versus MEDLINE as a whole:

$$\text{relative frequency} = \frac{f_{d,X,R}}{f_{d,X,M}}$$

In the replication of the first discovery, Gordon and Lindsay found that term frequency, document frequency, and the tf·idf statistic are strongly related. They used these first three statistics to select B-terms. Repeating the whole process, but with each of these B-terms, they used the fourth statistic, relative frequency, to discover the C-term. With this method, they managed to replicate Swanson's discovery. They hypothesized that this would be also be applicable to other discoveries.

In their attempt to replicate Swanson's second discovery this approach failed. The first three statistics did lead to discovering most of the B-topics also found by Swanson. Relative frequency, however, did not lead from one of these B's to the C-topic sought. Only when using a different method, where information from multiple B-terms was combined, the C-topic was discovered.

3.2.3 Weeber et al

Background

Marc Weeber and his colleagues developed a literature-based knowledge discover system called the DAD system. This system uses concepts to represent its topics. This system is described in (Weeber et al., 2000; Weeber et al., 2001). In the latter article it is also evaluated by simulating the first two discoveries of Swanson.

Using this system, Weeber managed to replicate Swanson's first two discoveries. In 2003, Weeber used his system to search for new therapeutic uses for the drug thalidomide. He found several diseases for which thalidomide might be helpful in (Weeber et al., 2003; Weeber, 2003).

Research

The DAD-system consists of both an open and a closed discovery process. The open process starts with a query describing topic A. This query is mapped to a concept using MetaMap. Synonyms are retrieved and lexical variants for these concepts are generated. These lexical variants are used for a PubMed query to retrieve all records containing topic A.

From these records, all sentences that contain topic A are selected. Using MetaMap, these sentences are mapped to concepts. All concepts that co-occur with topic A in a sentence are put in a list.

The list is filtered with a semantic filter. The user selects semantic categories of concepts that are likely to contain interesting B-concepts. Examples of such categories are 'diseases' and 'drugs'. The system then only retains those concepts fitting in one of these categories. The user chooses the B-concepts from the resulting list.

The described process is repeated with the selected B-concepts to get a list of C-concepts. Studying this list, the user can come up with one or more hypotheses of the existence of a relationship between A and some C.

These hypotheses can be strengthened by the closed process. In this process, records on C are retrieved, again using MetaMap. A list of concepts co-occurring with C in a sentence is retrieved. Next, only concepts occurring in both the C-list and the A-list are considered.

This B-list is reduced by applying the same semantic filters as used in the A-B step. The resulting list will probably highlight more potential pathways between A and C then was apparent in the open discovery process. The pathways can be verified by studying the sentences in which the A-B and B-C co-occurrence can be found.

3.2.4 Hristovski et al

Background

Another literature-based discovery system was developed by Dimitar Hristovski and his colleagues (Hristovski et al., 2001). This system uses MeSH descriptors (see section 2.3.1), because they incorporate human expert knowledge. According to Hristovski, this makes them better document descriptors

than title words. Association rules are used to discover relationships between topics. This system was evaluated by making potential discoveries in Medline publications with early publication dates and see how many of these discoveries become realised at later dates.

In (Hristovski et al., 2003), a new version of this system is described. The modified system is called BITOLA. The main difference with the earlier version system is the use of genetic knowledge. This is applicable when the system is used to find connections between a disease A and a gene C. The system uses domain knowledge about the chromosomal location of the candidate genes.

Research

In BITOLA, each MEDLINE record is represented by the MeSH descriptors assigned to the record and the gene symbols found in the title and abstract. The source of the gene symbols and names are a number of gene databases described in (Hristovski et al., 2003).

BITOLA starts with the calculation of all associations between MeSH descriptors in a part of Medline. The association rules used take the form of $X \rightarrow Y$ (confidence, support). Confidence is the percent of articles containing X which also contain Y. Support is the number of articles which contain both X and Y. The calculated associations are stored.

The system uses an open discovery process to discover new relations. Beginning with a topic of interest A, the system retrieves all MeSH descriptors for which $A \rightarrow B$ exists. Filtering can be done in two ways. The semantic filters also used by Weeber can be applied. Also, thresholds can be set on the support and confidence levels of the association rules.

The system then moves on from B to C in the same way as from A to B. Additional filtering of the C-list is done by excluding al C terms for which an $A \rightarrow C$ rule already exists. An optional filter requires that both A and C occur at the same chromosomal location, if that information is available. Ordering of the C-list can be done by confidence or support levels, or by semantic type.

3.2.5 Srinivasan

Background

In (Srinivasan, 2001), Padmini Srinivasan introduced a text mining tool that exploits the MeSH information accompanying MEDLINE records. It can be used to explore MeSH concepts and subheadings in a retrieved set of

documents. The tool has evolved since then and one of the functions offered now is that of concept exploration. This function can be used to build a profile for a concept from a text collection. A concept profile consists of a set of attributes that are strongly associated with the concept of interest in the text collection. (Srinivasan & Wedemeyer, 2003)

Srinivasan identified several possible uses for her concept exploration function. In (Srinivasan & Wedemeyer, 2003), she uses the function to study research trends over time. The differences between concept profiles of the same concept, but over a different time period, tell something about the changes in research over time. Another use of the exploration function was studied in (Srinivasan & Sehgal, 2003). Here, Srinivasan used her tool to identify similar drugs or genes, given a initial drug or gene concept. In (Srinivasan, 2004), a third use was studied. In this paper, Srinivasan, used her exploration function to build a literature-based discovery system fitting the framework described earlier. She used this system to successfully replicate many of the discoveries and hypotheses made by Swanson.

Research

The system developed by Srinivasan consists of both an open and a closed discovery process. As usual, the open process starts with a search in MEDLINE for documents about starting topic A. All MeSH terms extracted from this retrieved set R are candidate B-terms. These candidate B-terms are organized by MeSH semantic type. The list can be filtered by selecting semantic types. Only B-terms belonging to one or more of the selected types will be retained.

The B-terms are ranked within each semantic type by weight. The weights for each term are calculated using the *tf·idf* weighting scheme and are normalized within each semantic category. The weight of a B-term t_y in semantic type x ($t_{x,y}$) is:

$$w_{x,y} = \frac{v_{x,y}}{highest(v_{x,l})} \text{ with } l = 1, ..., m$$

where m is the total number of terms belonging to semantic type x. Further,

$$v_{x,y} = n_{x,y,R} \cdot \log \frac{N}{n_{x,y,M}}$$

where N is the number of documents on A retrieved from MEDLINE, $n_{x,y,M}$ is the number of documents in MEDLINE in which $t_{x,y}$ occurs and $n_{x,y,R}$ is the number of retrieved documents for A.

The B-terms are now further filtered by retaining only the top n (n is used defined) B-terms within each semantic type. The next step is a MEDLINE search for each of the B-terms left. From each of the retrieved sets, MeSH terms are retrieved. These can again be filtered by selecting semantic types. Organizing by semantic type and ranking by weight is also repeated. The lists are now combined, where the weight of each term is the sum of its weight in each of the separate lists. This combined list is filtered by excluding all terms for which a search in MEDLINE for documents containing both the A-terms and the candidate C-terms returns non zero results.

The result of the open process is a list of C-terms organized by semantic type and ranked within each semantic type.

The closed process starts with two MEDLINE searches for A and C. For both searches, a list of candidate B-terms is generated. These lists are organized by semantic types. They can be filtered by selecting semantic types. B-terms not falling in one of these types are excluded. The next step is the merging of these lists. Only items appearing in both lists are retained, and their new weight is the sum of their weights in the separate lists. The resulting list is filtered by excluding all items for which a search in MEDLINE for documents containing A, B, and C returns non zero results.

The result of the closed process is a list of B-terms organized by semantic type and ranked within each semantic type.

3.3 General issues in literature-based knowledge discovery

The systems above present different approaches to literature-based discovery. Although they all fit a general framework, they make different choices on several important points. By studying the choices made and the motivation behind those choices, we can form some ideas for our own system. We will compare the systems on three major issues, which are the knowledge source chosen (section 3.3.1), the knowledge representation chosen (section 3.3.2), and the choice between an open and/or a closed discovery process (section 3.3.3).

We will not provide a complete evaluation of these systems, to see which is the best, or to compare them to our own system. The difficulties involved with such a comparison become apparent when studying how the systems themselves were evaluated. This is discussed in section 3.3.4.

3.3.1 Knowledge source

In section 1.3, we argued that the scientific literature would be a good source of knowledge for a knowledge discovery system. Each of the systems discussed in this chapter use the scientific literature as a knowledge source. However, within the scientific literature, there are still choices to be made. There are different sources to derive the contents of a scientific article from. It is possible to use the full text of each article. Other possible sources are the title of the article, an abstract of it, or document descriptors such as MeSH descriptors. There are several differences among these four knowledge sources.

The first difference is the format in which the knowledge is represented. Full texts, abstracts and titles are in a *free text* format. The way humans read text, is difficult to simulate with a computer. Therefore, free text is harder to analyse automatically than structured forms of knowledge. Document descriptors are more structured, and thus easier to use with a computer.

The online availability of the sources is another difference. Titles and abstracts are readily available online (through PubMed). Full text only in a limited number of cases. Descriptors are only available for indexed articles.

Since document descriptors are assigned by human indexers, there is a time lag between the publication of an article and the availability of document descriptors. The other three sources also suffer a time lag, between their date of publication and their appearance in library systems. However, the time lag of document descriptors is much larger.

A fourth difference is the amount of text versus the focus of the text. Full text has more text (and thus contains more knowledge) than the other sources. Abstracts have less text, and titles even less than that. The amount of text used for descriptors vary. While full text has more text, it is less focussed than the others. Both abstracts and descriptors attempt to describe the main point of an article. The more focus a knowledge source has, the easier to extract the main point of an article from it. However, there is less knowledge to extract.

The choices made in the systems studied differ. Swanson uses titles and headings as knowledge sources. Gordon/Lindsay use complete MEDLINE records in their recent research. Weeber uses titles and abstracts. Hristovski and Srinivasan both use MeSH descriptors.

We would like to build a system that is independent of human indexing, to make it as general as possible and enable it to use knowledge with a minimal time lag. Therefore, our system cannot depend on just document

descriptors. However, we will include them when available. Since the other sources are in the free text format, our system must be able to deal with free text. As full texts are not readily available online, we will not use them as a knowledge source. This leaves titles and abstracts. Titles are even more focussed than abstracts, but the amount of knowledge in them is limited. However, since both are available and in the same format, we will use both for our system.

To sum up, we will use titles, abstracts, and document descriptors for our system. Since we will apply our system to the biomedical literature, we will use complete MEDLINE records (which contain all three).

3.3.2 Knowledge representation

Another important choice to be made is how to represent the knowledge contained in articles. Closely related is the question of how to determine which topics are contained in a certain document. Swanson and Gordon/Lindsay used words or short phrases to represent topics. Weeber chose to use UMLS concepts; Hritovski and Srinivasan chose MeSH concepts. Note that the MeSH concept set is a subset of the UMLS set.

Words and short phrases are easily extracted from free text. When a word or phrase occurs in a piece of text, the topic represented occurs in that piece of text. The problem is determining which are meaningful words or phrases. Swanson and Gordon/Lindsay use stop lists to exclude meaningless words. Stop lists, however, have to be domain specific to be very useful and would therefore have to be build anew for each new discovery.

The advantage of using concepts, like Weeber, Hritovski, and Srinivasan do, are manyfold. Using concepts ensures that all candidates for topics retrieved from a document are meaningful. This is especially useful in the case of multiple word phrases. Using a domain-specific thesaurus ensures that all concepts retrieved are relevant to that domain. The two advantages above lead to a third: there is no need for a user-defined stop list. A fourth advantage is the collapsing of synonyms and textual variants in a single concept. This will ensure that topics with many synonyms and/or textual variants will not go unnoticed. A fifth advantage is the possibility to exploit the semantic types and relationships often associated with thesauri.

The disadvantages of using concepts in combination with a representation of source literature which is in the free text format include the difficulties involved with extracting concepts from free text. Hritovski and Srinivasan didn't have this problem, the MeSH descriptors they use to represent an instance if literature are all MeSH concepts. Weeber uses MetaMap to extract

concepts from the free text. Other disadvantages of using concepts surface when we use a domain-specific thesaurus or when we exploit the semantic types and relationships often associated with thesauri. Both will limit the domains where the system can be applied to the ones where these resources are available.

Because we choose to use a knowledge source which is partly in free text format, we will be faced with the challenge of automatically extracting concepts from free text if we choose to use concepts to represent knowledge. The Collexis software described in section 2.3.3 provides a solution for this. A lot of semantic information is lost when using fingerprints instead of text, but we hope that the advantages of using concepts outweigh this.

3.3.3 Open and/or closed discovery process

We think that both the open and the closed process contribute to the discovery process. The open process is the most important, because it generates hypotheses, which is the goal of the discovery system as a whole. However, the closed process can supply evidence to support or weaken the generated hypotheses, and plays an important role in determining whether to further pursue the discovery.

Both Gordon/Lindsay and Hritovski focus on an open process. Swanson, Weeber, and Srinivasan use both processes. We think our system should include both an open and a closed discovery process. However, this thesis will only cover the open process.

3.3.4 Evaluating systems

The goal of the literature-based discovery systems described in this chapter is to discover unknown, but valid relationships between topics described in literature. The systems should therefore be evaluated by their ability to discover unknown, but valid relationships. It is feasible to check if the relationships 'discovered' by the systems are unknown. You can search the scientific literature to see if any of the relationships are directly described. The scope of the available literature and the difficulties involved with determining automatically what knowledge a document contains make this difficult. However, as Swanson shows in his articles, it is feasible to ensure that every discovered relationship is at least relatively unknown.

To see if the discovered relationships are valid is much more difficult. We can use a closed discovery process to strengthen each hypothesis of the existence of a relationship. However, eventually the hypotheses will have to

be judged by human experts. While these experts may be able to dismiss several relationships off hand, much experimentation may be required to verify even one of the discovered relationships. Because of this, we will need another evaluation method than simply try to discover unknown relationships.

All but one of the systems discussed in this chapter are evaluated in the same way, which is trying to replicate discoveries made by Swanson. Swanson's discoveries are well documented, and a couple of them have been supported by (experimental) evidence. Swanson himself evaluates his ARROWSMITH tool by trying to replicate his *Raynaud's disease and fish oil*, his *Migraine and Magnesium*, and his *Somatomedin C and Argine* discoveries. Both Gordon/Lindsay and Weeber use the first two of these discoveries to evaluate their systems. Srinivasan uses not only all three, but also his *Indomethacin and Alzheimer's disease* and *Calcium-Independent Phospholipase A2 and Schizophrenia* discoveries.

The methods used by the different researchers to evaluate the systems on these cases are very similar. They all start their discovery systems with one end of the relationship, and see if they end up at the other end. The criteria used is if they can argue that the connection would be made with a reasonable amount of effort by an expert user.

Hristovski used another evaluation method. He used his system to discover relationships in a body of literature published before a certain date. The relationships must be unknown at that date. Next, he checks which new relationships have been discovered since the date. He does this by searching the literature published since that date for documents where two concepts appear together which did not appear together in the literature from before the date. His system is judged by checking the percentage of the new published relationships which have been 'predicted' by his system and the percentage of his predictions which have been realised.

Both these methods of evaluation have their own problems when used for comparing the systems. Successfully replicating some cases by Swanson only shows that the systems are indeed capable of discovering some relationships. Success or failure in this does not say much of the capability of the systems to discover other relationships. Hristovski's method may be more useful, but can only be used to compare completely automated systems. Most of the other systems rely on human input to make discoveries. The effort involved in making enough discoveries to do a good comparison is huge and certainly beyond the scope of this thesis. We will use the first method to evaluate our own system.

Chapter 4

System description

In this chapter, we present a new system for literature-based knowledge discovery. We use the general architecture and the discussed issues in the previous chapter as a guideline. Thus, we will use complete MEDLINE records as our knowledge source. Topics will be represented by concepts. Also, the system should allow for both an open and a closed discovery process. However, this thesis will only cover the open discovery process.

A part of the architecture not discussed so far is the filtering and sorting of candidate concepts (B's and C's). This part of a literature-based discovery system is both the most important and most difficult one. An enormous list of words, terms or concepts related with the *seed concept* (A) is no more useful to a scientist than the complete collection of articles including A. The system should filter and/or sort this list to put forward the topics that are most relevant to the scientist. For this filtering/sorting step, we will use the tool described in the next section. We used it in two different approaches, which are described in section 4.2 and section 4.3. We suggest some modifications that may improve the results of our two basic approaches in section 4.4.

4.1 Associative Concept Space

The Associative Concept Space (ACS) is a n-dimensional space in which concepts are positioned. Along with the positions of each concept, the ACS stores the connections among the concepts. A connection between two concepts reflects their co-occurrence in one or more articles. The position of a concept reflects the connections with the other concepts in the ACS. Concepts that have many connections, being either direct or through interme-

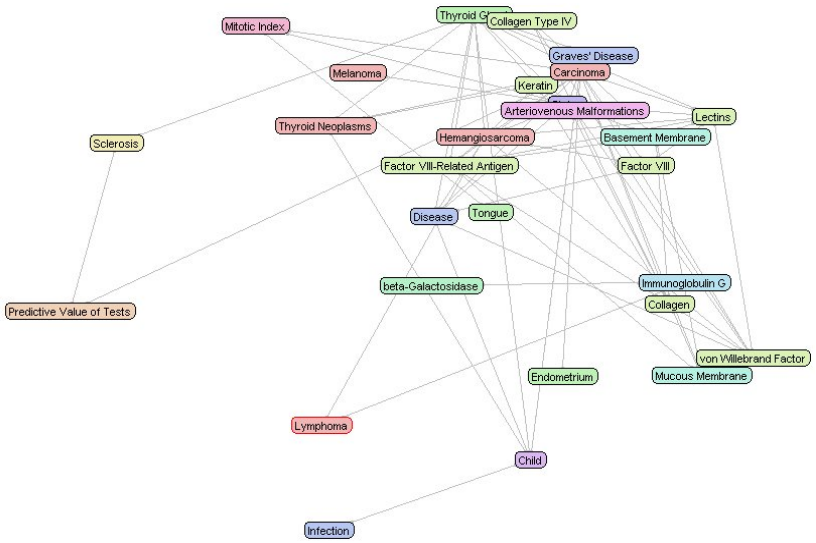

Figure 4.1: A 2-dimensional projection of an ACS. Concepts are represented by their names. Connections among concepts are represented by lines. (Biosemantics Group Rotterdam, 2003)

diate concepts, are positioned closer to each other than concepts with fewer connections.

The algorithm for constructing such an ACS was introduced in (Schuemie, 1998; Schuemie & van den Berg, 1998; van den Berg & Schuemie, 1999; Schuemie & van den Berg, 1999) and further developed in (van der Eijk, 2001; van der Eijk et al., 2002; van Mulligen et al., 2002; van der Eijk et al., 2004). A 2-dimensional projection of a small ACS is shown in figure 4.1.

An ACS is constructed from a set of Collexis fingerprints. As described in chapter 2, a Collexis fingerprint consists of a list of concepts. Each concept in a fingerprint has a weight associated with it, reflecting the importance of that concept in the text. Because the low-weight concepts are less likely to be important in the text represented by the fingerprint, the ACS only uses concepts for which the weight falls above a certain threshold.

Each remaining concept c_i in the fingerprint set L gets a n-dimensional location vector x_i in the ACS with randomly assigned coordinates:

$$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$$

After this, two rules are applied each learning cycle t. The first rule, the *learning rule*, moves all the concepts in each fingerprint to the *centroid* of that fingerprint. The centroid of a fingerprint f_k with m concepts (p_k) is defined as the average of the concept vectors $x_h (h = 1, 2, ...m)$ in f_k:

$$p_k = [p_{k,1}, \ldots, p_{k,n}] = \left[\frac{\sum_{h=1}^{m} x_{h,1}}{m}, \ldots, \frac{\sum_{h=1}^{m} x_{h,n}}{m} \right]$$

The learning rule is:

$$\forall i : x_i(t+1) = x_i(t) + \eta(t) \frac{p_k(t) - x_i(t)}{\|p_k(t) - x_i(t)\|}$$

where $\eta(t)$ is the *learning rate*. The learning rate is defined as:

$$\eta(t) = \frac{2}{\min(t, u)}$$

with u a constant set by the user.

After the learning rule is applied for each fingerprint in L the second rule, the *forgetting rule*, is applied. This rule moves all concepts in L away from the centroid p_L, which is defined as the average of the vectors of all the concepts in L. This separates the concepts and prevents congregation of the concepts in one point. The forgetting rule is:

$$\forall i : x_i(t+1) = x_i(t) - \lambda(\|p_L(t) - x_i(t)\|) \frac{p_L(t) - x_i(t)}{\|p_L(t) - x_i(t)\|}$$

where $\lambda(x)$ is defined as:

$$\lambda(x) = \begin{cases} 1 & \text{for } x < 1 \\ 1/x & \text{for } x \geq 1 \end{cases}$$

After these steps have been repeated an user defined number of cycles T, the ACS is trained. It can then be used for discovery purposes, such as retrieving the k closest concepts to a given seed concept. For determining the distance between two concepts i and j, the Euclidean distance between their vectors ($d_{i,j}$) is used:

$$d_{i,j} = \sqrt{\sum_{l=1}^{n} (x_{i,l} - x_{j,l})^2}$$

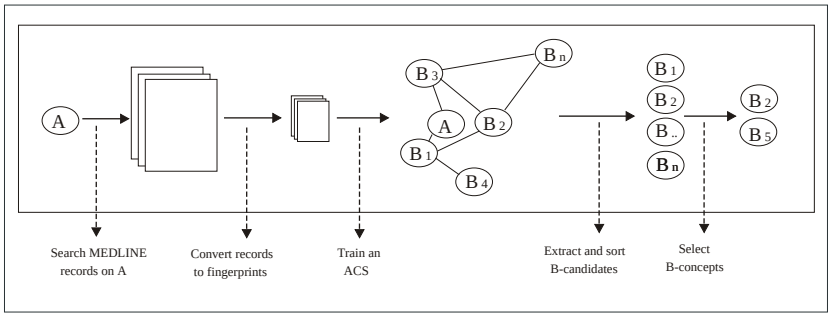

Figure 4.2: The $A - B$ step of the two-step approach.

4.2 Two-step approach

The first of our two approaches closely fits the general architecture for an open discovery process described in chapter 3. Since it consists of an $A - B$ step and a $B - C$ step, we call it the *two-step approach*.

Step one: the $A - B$ step

The first step starts with the downloading of MEDLINE records on the concept of interest, A. These records are converted into fingerprints using the Collexis software described in 2.3.3. With the resulting set of fingerprints, an ACS is trained.

The concepts contained in this ACS are A and candidates for B-concepts. Note that this list of candidates has already been filtered by the cutoff used in the training of the ACS. Only concepts with a weight above a given threshold are included in the ACS. Further filtering is not done, but the candidates are ranked to put the most interesting forward. For this, the distance in the ACS between the candidate B-concepts and A is used. Due to the training process, concepts close to A are supposed to have a stronger relationship with it than concepts that are further away.

From the ranked list of B-concepts, the user selects one or more to complete the $A - B$ step. The $A - B$ step is illustrated in figure 4.2. If the user selects more than one B-concept, the $A - B$ step is followed by multiple $B - C$ steps, one for each B. In this way, multiple interesting $A - B$ combinations can be explored.

Step two: the $B - C$ step

In the second step, MEDLINE records on the B-concept selected in the $A - B$ step are downloaded and converted to fingerprints. The resulting set of fingerprints is filtered by excluding the fingerprints which contain A. An ACS is trained on the filtered set.

The concepts in this ACS are candidates for C-concepts. These concepts are ranked by distance from B and the user selects one or more C-concepts. Together with A, these form hypotheses of relationships between A and some C, which can then be used for a closed discovery process.

Finally

The approach introduced above not only uses direct co-occurrence to determine the strength of a relationship, but might also exploit indirect co-occurrence through other concepts. If two concepts are strongly related, there should be many direct and indirect co-occurrences in the literature. The ACS can be used to exploit this information and, by combining an $A - B$ and $B - C$ relationship, discover strong and novel relationships between previously unconnected concepts.

4.3 One-step approach

Our second approach skips the B-concepts and makes one step from A to C. This approach requires some pre-processing before the actual discovery process starts.

Pre-processing

Before discoveries can be made, an ACS is trained on a large body of scientific literature. This should cover as much literature of interest for potential discoveries as currently feasible. Examples include an entire year of articles published in MEDLINE, or all articles that appeared in a certain journal. The ACS can be used to make discoveries with $A - C$ steps.

The $A - C$ step

A seed concept A can be selected from the ACS. From the ACS, a list of candidate C-concepts is extracted. These are the concepts that do not co-occur with A. The list is sorted ascending by distance between A and each C-candidate. From this ranked candidate list, C-concepts can be selected

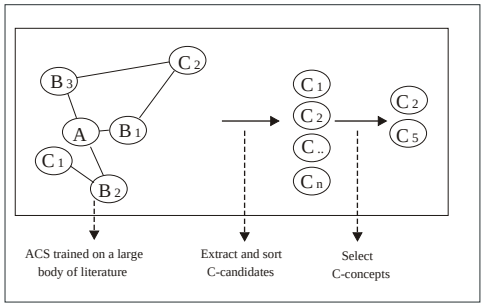

Figure 4.3: The one-step approach.

and used for the following closed process. The $A - C$ step is illustrated in figure 4.3.

Finally

The one-step approach uses less input from the user than the two-step approach, because no B-concepts are selected. It does possibly exploit more indirect information, because the ACS is trained on a larger and more general corpus of literature. In such a corpus, there are more documents and thus more possibilities for indirect relationships between A and C.

4.4 Modifications

In addition to the two basic approaches presented above, we present four possible modifications to these approaches here. They may improve the results of our basic approaches, and will be evaluated along with them.

Combining rankings

In our basic two-step approach, candidate concepts are ranked by distance to the seed concept. There are several more ways to rank these concepts based on information available in the fingerprints or the ACS. It is possible that these other rankings contain other information of the relationship between seed and candidates than the ranking based on distance. In this first modification to the basic two-step approach, we seek to exploit this information.

Additional to the ranking ascending based on *distance* (D), which is used in the basic approach, we discern six ways to rank candidate concepts:

- Descending based on *co-occurrence* (C), the number of fingerprints which include both the seed concept and the candidate concept.
- Descending based on *average seed weight* (S), which measures the average weight of the seed concept in the fingerprints which include both the seed concept and the candidate concept.
- Descending based on *average candidate weight* (W), the average weight of the candidate concept in the fingerprints which include both the seed concept and the candidate concept.
- Descending based on WS ($W \cdot S$).
- Descending based on CW ($C \cdot W$).
- Descending based on CWS ($C \cdot W \cdot S$).

The last two ways of ranking presented above can be seen as examples of *weighted co-occurrence*, because they use not only information of the number of co-occurrences, but also of the importance of each co-occurrence, which is reflected by the weight of the candidate and/or seed.

Some of the information expressed by the different ways of ranking is already used in the training of the ACS and is thus reflected in the ranking based on distance. The information provided by the average candidate weight is used to exclude low-weight concepts from ACS training. Co-occurrence information is at the core of ACS training. However, not all information in the average candidate weight is used, and the information which is used might not be (fully) reflected by ACS distance. Average seed weight is not used at all in ACS training.

We try to exploit the information contained in rankings not based on distance by combining other rankings with the ranking based on distance. To combine two or more rankings, the positions of candidate concepts in them are averaged. A new ranking is then computed based on these average values. For example, when a candidate x has position 3 in one ranking, and position 5 in another (average 4), while a candidate y has positions 1 and 11 (average 6), candidate x will be ranked higher in the combined ranking.

Using a more general document set

The main strength of the ACS is the use of indirect co-occurrence information. Concepts do not only appear closer together if they co-occur much in the training set, but also if there are a lot of connections through (one or more) intermediate concepts. In our basic two-step approach, we train each ACS on a document set obtained by searching for documents on the seed concept. This means that the set includes all documents mentioning both the seed concept and one or more candidate concepts. Thus, we use all direct co-occurrence information available in the training of the ACS. However, this set will probably not include all indirect co-occurrence information.

For example, a candidate concept x has several strong indirect relationships with the seed concept s through another concept y. The documents describing the relationship between s and y are in the document set, since that set contains all documents on y. There are also some documents describing a (weak) direct relationship between s and x (otherwise it would not be a candidate concept). However, the documents describing the relationship between y and x are not in the document set. In this example, s and x are positioned far from each other in the ACS, based on the weak direct relationship. When the documents describing the relationship between y and x would be included in the training set of the ACS, they would be positioned closer to each other.

The idea presented here is identical to the one underlying the use of a large and general document set for the one-step approach. The difference is the use of user selected B-concepts in this approach, which does not happen in the one-step approach. We will explore the effects of using more indirect co-occurrence information by training the ACS using a document set based on a query that is more general than that used in the basic approach, thus including more indirect information in the set.

Using inverse document frequency

In both our approaches, the rankings are based on the distance between seed concept and candidate concepts in the ACS. The position of concepts in the ACS is influenced by their co-occurrence with other concepts. Two concepts co-occur if they appear in the same fingerprint. However, for the training of the ACS, a cutoff is used. All concepts in each fingerprint with a weight below a certain cutoff value, are excluded from that fingerprint. The weights are influenced by six statistics, which represent the importance of

that concept in the source document of the fingerprint (see section 2.3.3). This way, concepts who are unimportant in the document, do not influence ACS training.

The weights do not reflect the generality of the concepts. General concepts have the same influence on ACS training as more specific concepts. However, the presence of a general concept in a fingerprint had less significance than the appearance of a specific concept. Very general concepts, such as 'Human', appear in many documents. They are thus likely to be present in a randomly chosen fingerprint, and their presence is not very informative. The presence of a more specific concept, which is not likely to be present in a randomly chosen fingerprint, is more informative. A way to improve ACS training, and thus our rankings, would be to reflect the generality of concepts in the fingerprint weights. More general concepts should receive less weight, and less general concepts more.

We will explore the effects of compensating for generality by using *inverse document frequency (IDF)* to correct the fingerprints used for ACS training. In each fingerprint, the new weight $w_{i,j}$ of concept c_i in fingerprint f_j is:

$$w_{i,j} = v_{i,j}/highest(v_{i,l}) \text{ with } l = 1, ..., m$$

where m is the total number of concepts belonging to fingerprint f_j. Further,

$$v_{i,j} = u_{i,j} \cdot (\log \frac{N}{f_{d,X,M}} + 1)$$

where $u_{i,j}$ is the old weight of concept c_i in fingerprint f_j, N is the number of documents in a large document set M, and $f_{d,X,M}$ is the number of documents in M in which c_i is present.

This is similar to the compensating Gordon/Lindsay and Srinivasan did in their systems (see section 3.2.2 and section 3.2.5).

Using semantic categories

In both our approaches, candidate concepts are presented in a long, ranked list. This list contains all concepts related to a seed concept. Some of these concepts might be very different. For example, 'blood viscosity' and 'Japan'. This makes it harder for a user to see which concepts are important in the discovery process. He might be more interested to see how concepts in a certain category are ranked. For example, which 'chemical' is most strongly related to the seed concept. When we present the candidate concepts in categories, the user might notice relationships more easily. Also, the user can apply filtering by ignoring concepts in certain categories. This last

Table 4.1: Semantic categories

Activities & Behaviors	Disorders	Objects
Anatomy	Drugs	Occupations
Chemicals	Genes & Molecular Sequences	Phenomena
Concepts	Ideas	Physiology
Devices	Living Beings	Procedures

possibility has been used by both Weeber and Srinivasan in their discovery systems (see section 3.2.3 and section 3.2.5).

We will use the semantic types assigned to concepts in the UMLS semantic network (section 2.3.2). There are 134 semantic types in this network, which have been aggregated in 15 semantic categories in (McCray et al., 2001). These 15 categories are listed in table 4.1. Each contains several UMLS semantic types. For example, the semantic category 'Anatomy' contains, among others, the semantic types 'Body Location or Region', 'Cell', and 'Tissue'. A semantic type 'Body Location or Region' contains such concepts as 'Arm', 'Toes', and 'Leg'.

We will explore the use of semantic categories by presenting candidate B or C candidates divided among these 15 categories. We can then use the possibility to focus on one or more categories during the discovery process.

Chapter 5

Method of evaluation

This chapter discusses how we evaluate the two approaches to a literature-based knowledge discovery system presented in chapter 4. We discern three objectives of this evaluation:

1. To evaluate the added value of the ACS in the discovery process.
2. To evaluate whether any of the possible modifications improve the basic discovery process and thus should be used.
3. To evaluate whether a scientist using either the two-step or the one-step approach will be able to make discoveries with a reasonable amount of effort.

We discussed evaluating literature-based discovery systems in 3.3.4. We will evaluate our system by simulating a former discovery, a test case. The discovery we will use for this is the first discovery made by Swanson, which is described further in section 5.1. To meet our first two evaluation objectives, we want to compare different rankings of candidate concepts. We will use a technique called *ROC analysis* for this. This technique is further discussed in 5.2. The methodology used in our evaluation is described in section 5.3.

5.1 Test case

There are a couple of well-documented examples of literature-based discovery. One of them is the already mentioned first discovery of Swanson. We will use this discovery as a test case for our evaluation. We first provide a detailed description and then discuss how we use this test case in our evaluation.

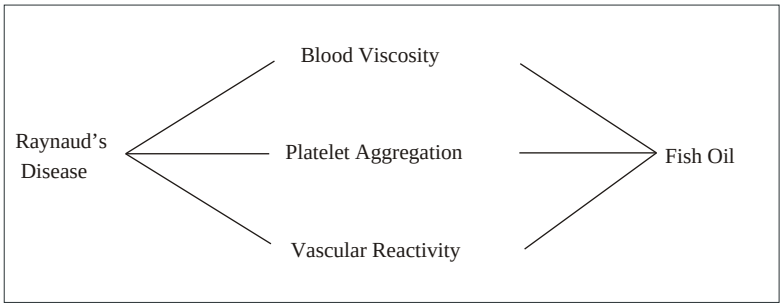

Figure 5.1: The three pathways through which Raynaud's Disease is connected to fish oil.

5.1.1 Description

Swanson's first discovery was a relationship between *Raynaud's disease* and *fish oil*. Raynaud's disease is a condition that causes some areas of the body, such as fingers, toes, tip of the nose, and ears, to feel numb and cool in response to cold temperatures or stress. It is a disorder of the blood vessels that supply blood to the skin. During a Raynaud's attack, there is limited blood circulation to affected areas.

Fish oil is present in high amounts in some fish, particularly fatty types prevalent in cold water, such as salmon, mackerel, and herring. It has been shown to improve blood circulation.

Swanson hypothesized that dietary fish oil might ameliorate or prevent Raynaud's disease. The relationship was not described directly in the medical literature, but Swanson found three indirect connections. Dietary fish oil had been shown to reduce *platelet aggregability* and *blood viscosity*. It also indirectly causes *vasodilation*. Patients with Raynaud's disease have intermittent blood flow in their extremities. This is caused by high platelet aggregability, high blood viscosity, and vasoconstriction. These three pathways are shown in figure 5.1.

The effects of fish oil, and especially Eicosapentaenoic acid (EPA), which is a fatty acid found in fish, were first discovered in an Eskimo population. Their low levels of blood cholesterol, triglycerides, and low-density lipoproteins and the low incidence of myocardial infarctions in the population was found to be a result of their EPA-rich diet.

Later EPA-rich experimental diets have been reported to reduce platelet

aggregation. EPA leads to the synthesis of the prostaglandin PG13, which strongly reduces platelet aggregation. There are also other mechanisms through which EPA reduces platelet aggregation.

The vasodilating effects of fish oils have been reported to occur in rats. There are also several other reasons to suspect this effect. PG13 (produced by EPA) can be presumed to be a strong vasodilator because of its similarity to PG12. Furthermore, EPA's effect on platelet aggregation suppresses vasoconstriction, since aggregating platelets normally release vasoconstricting substances.

The third effect of fish oil is that it reduces blood viscosity. Several laboratories have found this effect. EPA incorporates into cell-membrane phospholipids, which results in improved erythrocyte fluidity or deformability, which in turn leads to viscosity reduction. Fish oil also reduces blood lipids, especially triglycerides. The blood levels of triglycerides have been shown to be directly related to blood viscosity.

In patients with Raynaud's disease, abnormally high platelet aggregability, high blood triglycerides, and high blood viscosity have been reported. Both Prostaglandin E1 and prostacyclin have been successful in treating Raynaud's disease, which is thought to be because of their effect on platelet aggregation and their vasodilating effect. Another vasodilator, Nifedipine, has also been successful against Raynaud's. Finally, a selective antagonist of the serotonin receptor, Ketanserin, has had some success in treating Raynaud through reducing blood viscosity (Swanson, 1986).

The hypothesis following from this, that dietary fish oil might ameliorate or prevent Raynaud's disease, has later been shown to exist in a clinical trial (Chang et al., 1988; DiGiacomo et al., 1989). This, and the fact that is well documented by Swanson, makes it a good case to evaluate our algorithms on.

5.1.2 Usage in the evaluation

We will use our system to simulate the discovery made by Swanson. A scientist using our system to do a discovery will have to spend time when selecting B or C candidates from the ranked lists presented by our system. The higher B or C candidates *relevant* to the discovery process are ranked, the less time the user will have to spend to discover them. For the second and third evaluation objectives we will compare different rankings using ROC-analysis. The ROC-analysis is based on the ranking of relevant candidates versus that of non-relevant concepts.

So, for all three of our evaluation objectives, we need to assess the quality

of rankings. This quality is determined by the position of relevant concepts in the ranking. So far, we haven't discussed what relevant concepts are. Relevant B-candidates are those that will lead to a C-concept. A relevant C-concept is one that has an unknown, but valid relationship with A. Since we use the Raynaud's disease case for our evaluation, relevant B-candidates are those that will lead to fish oil. The relevant C-candidates are fish oil and concepts very similar to the concept of fish oil.

Relevant B-concepts

We evaluate our two-step approach by studying the ranking of B-candidates. To be able to do our evaluation, we used our knowledge of the Raynaud's Disease case to label these candidates as *relevant* or *not relevant*. Within the relevant group, we discern *highly relevant* concepts and *not highly relevant* concepts. Those concepts considered highly relevant are very similar to the B-concepts Swanson describes. Those that are not highly relevant are less similar. We list the B-concepts we considered highly relevant in table 5.1. A table listing those we considered not highly relevant can be found in the appendix. Both tables are divided into the three pathways discerned by Swanson. This is for presentation purposes only, our evaluation method did not use this distinction.

Relevant C-concepts

In the second step of our two-step approach and in the one-step approach, we need to assess the quality of rankings of C-concepts. For this, we labelled C-candidates *relevant* or *not relevant*. The relevant C-concepts are listed in table 5.2.

5.2 ROC analysis

We need an algorithm that measures the quality of a ranking. For example, the ranking of B-candidates which co-occur with the seed concept A. Some of the B-candidates are relevant to our discovery process and should be used as B-concepts to continue the search. We would like to rank the relevant ones as high as possible, to bring them to the attention of the user of the system.

To judge a certain ranking of concepts, we will look at the area that a scientist using our system will study and use to select concepts from. (In the B-candidates case, B-concepts to search further with.) The system should

Table 5.1: Highly relevant B-concepts

	Blood Viscocity	
Blood Circulation	Blood Flow Velocity	Blood Viscosity
Hemodilution	Viscosity	
	Platelet Aggregation	
Agglutinins	Alprostadil	Anticoagulants
Antithrombins	beta-Thromboglobulin	Blood Coagulation
Blood Coagulation Disorders	Blood Coagulation Tests	Blood Platelets
Dipyridamole	Edetic Acid	Epoprostenol
Erythrocyte Aggregation	Erythrocyte Deformability	Fibrin
Fibrinolysis	Fibrinolytic Agents	Platelet Activation
Platelet Adhesiveness	Platelet Aggregation	Thrombosis
Thromboxanes	Venous Thrombosis	
	Vascular Reactivity	
Alprostadil	Blood Circulation	Blood Flow Velocity
Blood Pressure	Capillary Permeability	Capillary Resistance
Dipyridamole	Epoprostenol	Vascular Resistance
Vasoconstriction	Vasodilation	

Table 5.2: Relevant C-concepts

Fatty Acids, Essential
Fish Oils
Cod Liver Oil
Docosahexaenoic Acids
Fatty Acids, Omega-3
5,8,11,14,17-Eicosapentaenoic Acid

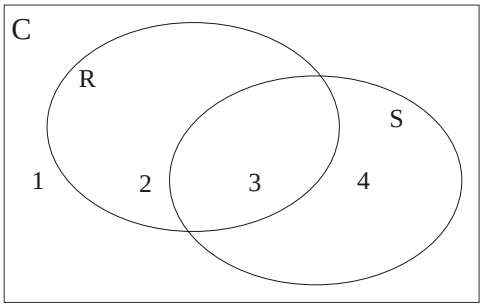

Figure 5.2: ROC sets. C is the set of all candidate concepts, R the set of relevant concepts and S the selection of the n highest ranking concepts. Area 3 should be as large as possible, while keeping area 4 small.

rank relevant concepts higher than non-relevant concepts. Thus, a scientist should find the relevant concepts R in the n highest ranking concepts. We will call this selection S. It is a subset of the complete collection of candidate concepts, as is R. This situation is illustrated in figure 5.2.

The selection S should cover as much of R as possible (area 3 should be large), because the scientist using the system will have more chance to make a discovery when the system presents him all concepts relevant to that discovery. This could be accomplished by making S large enough to include the whole of R. However, S would then also include more non-relevant concepts (area 4 would be larger). The user has to spend unnecessary time and effort to study and dismiss the concepts in area 4, making the discovery process harder.

There is thus a trade-off between the number of relevant concepts presented to the user and the time and effort the user has to spend studying the presented concepts. This trade-off is influenced by two things. The size of S, which is determined by n, and the position of S, which is determined by the quality of the ranking of the concepts in C. (A good ranking would position S over R.) For a given n, we would judge a ranking based on the number of relevant concepts in S. However, since n depends on the number of relevant concepts present and on the time and effort the user is willing to spend, we need a way to judge rankings based on a large number of possible values of n.

A way to do this comes in the form of Receiver Operating Character-

istic (ROC) analysis. We will adopt the ROC analysis described in (Metz, 1978) for our purposes. In this adopted algorithm, the fraction of all relevant concepts present in the top n concepts is defined as the True Relevant Fraction:

$$\text{TRF} = \frac{|R \cap S|}{|R|}$$

Similar, the fraction of all not relevant concepts that are present in the selection, the False Relevant Fraction:

$$\text{FRF} = \frac{|\neg R \cap S|}{|\neg R|}$$

There are two similar figures for concepts not included in the selection. The True Non-relevant Fraction:

$$\text{TNF} = \frac{|\neg R \cap \neg S|}{|\neg R|}$$

And finally, the False Non-relevant Fraction:

$$\text{FNF} = \frac{|R \cap \neg S|}{|R|}$$

Of these four, we are most interested in the first two, the TRF and the FRF. The TRF reflects area 3, and should be as high as possible, capturing as many of the relevant concepts possible in the selection. The FRF reflects area 4, and should be as low as possible, keeping the number of not relevant concepts in the selection as low as possible. As n increases from 0 to the total number of concepts, both fractions will increase from 0 to 1. Any value chosen for n will mean a trade-off between the two fractions.

Different values of n will result in different pairs of FRF and TRF. We can plot these pairs as x and y coordinate values of points on a graph. The points representing all possible combinations of FRF and TRF form a curve. This curve is the ROC curve. An example ROC curve is plotted in figure 5.3. A ROC curve always passes the lower left corner (n is 0), where we select no concepts, and thus include no not relevant concepts in our selection, but also miss all relevant concepts. The curve also always passes the upper right corner, where we include all concepts in our selection, thus including all relevant concepts, but also all non-relevant ones. A ROC curve of a ranking should be above the lower left to upper right diagonal of the ROC space, because we otherwise would be better off by looking at the concepts not selected ($\neg S$) by our system.

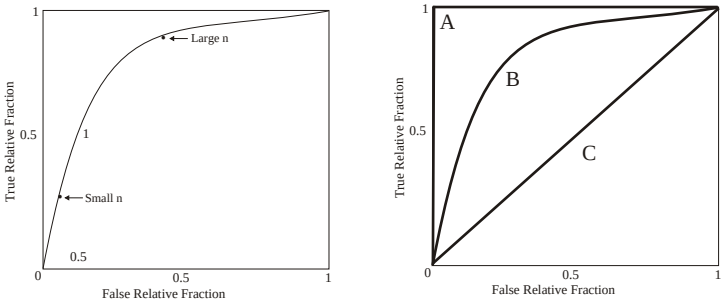

Figure 5.3: Examples of ROC curves. The left graph shows a typical ROC curve. The small n point shows a low FRF and low TRF. A larger n increases both values. The right graph shows three different ROC-curves. A is a ROC curve of a perfect ranking, B a more typical curve, and C a ROC curve of a ranking where the concepts have random positions.

If we obtain ROC curves for each different way of ranking, the rankings can be compared. In general, a ranking with a ROC curve more to the upper left is better. A numeric value that reflects this is the *Area Under the Curve* (AUC). It is the fraction of the area of the graph that falls under the ROC curve (Ming, 2002). An interesting property of the AUC is that its value is equal to the chance that a relevant concept will be ranked higher than a non-relevant one (Fawcett, 2003). The AUC of a ranking i is:

$$AUC_i = \Pr(Pos(c_r, i) > Pos(c_{\neg r}, i))$$

where $Pos(c_r, i)$ is the position of a randomly chosen relevant concept c_r in i and $Pos(c_{\neg r}, i)$ is the position of a randomly chosen non-relevant concept $c_{\neg r}$ in i.

We will use the AUC value to compare different rankings. It represent the quality of a ranking for all possible values of n. In a literature-based discovery system, the number of candidate concepts is often quite large. Because of this, we might only be interested in the quality of rankings for small values of n (making the assumption that the user will never study more than, for example, two fifth of all candidates). To be able to compare rankings for these smaller values of n, we will also calculate an AUC for these cases. (See figure 5.4.) In fact, we will calculate the AUC with n ranging from 0 to one fifth of the total number of concepts, ranging to two fifth, to

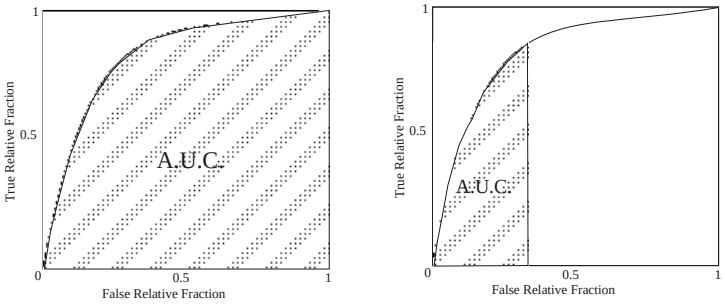

Figure 5.4: Area Under the Curve. The left graph shows the AUC of an example ROC curve. The right graph shows the AUC measured over a range of small values of n.

three fifth, to four fifth, and finally ranging over all the cases. These values for smaller values of n are limited. They can not be used to assess chances, as the complete AUC does. They can only be used for the comparison of different rankings.

5.3 Evaluation methodology

This section discusses the methodology we used to evaluate our system. We begin with the two-step approach and its modifications in section 5.3.1 and the one-step approach and its modifications in section 5.3.2.

5.3.1 Two-step approach

This process starts with downloading articles on A from MEDLINE. The resulting set of documents is converted into a set of fingerprints by Collexis software. Using this set of fingerprints, an ACS is trained. The concepts in this ACS are the candidates for B-concepts in an A to B, B to C discovery process. A ranked and filtered list of these candidates is extracted from the ACS.

Basic approach

For the basic approach, we did a search in MEDLINE with the following query:

Table 5.3: Default ACS parameter values

number of dimensions	n	8
constant for learning rate	u	10
number of learning cycles	T	10
fingerprint cutoff	$w_{\min}$	0.4

Table 5.4: Different rankings used in experiment 1-2.

D	C	W	S
CW	CWS	WS	D & C
D & S	D & W	C & S	C & W
S & W	D & C & S	D & C & W	D & S & W
C & S & W	D & C & S & W	D & CSW	D & CW

```
raynaud s disease AND 1900:1985[dp]
```

We used the fingerprints made from these documents were used to train an ACS. Training the ACS was done with the default parameters, which were adopted from (van der Eijk et al., 2004) and are listed in table 5.3. We sorted the B-candidates in this ACS ascending by their distance from the seed concept. To evaluate the use of the ACS, we also sorted them descending by co-occurrence with the seed concept.

Combining rankings

To evaluate the use of combining ranking, we tested a total of 20 rankings. They are listed in table 5.4. The others are combinations of rankings. These methods of ranking B-candidates were applied to candidates from the same ACS used in the basic approach.

Using a more general document set

To see the effect of a larger ACS on the ranking based on distance, we used a more general query to obtain our initial document set. The query used was:

```
vascular diseases AND (peripheral OR extremities OR finger OR
fingers OR toe OR toes) AND 1900:1985[dp]
```

The ACS trained on the resulting document set was trained using the same parameters as in the basic approach.

Using inverse document frequency

To test the effect of the use of IDF on our rankings, we trained an ACS based on a corrected version of the fingerprint set used in the basic approach. The fingerprints are corrected using IDF. The IDF values are calculated using a set of Collexis fingerprints constructed using all MEDLINE records with an entry date between the year 1996 and the year 2000. The compensated fingerprints are used to train an ACS, with the default parameters. Note that because the weights are corrected, different concepts are cut from the fingerprints. This affects both distance and co-occurrence.

Using semantic categories

For this last modification of the two-step approach, we used the same ACS used in the basic approach. We then divided the B-candidates over the 15 groups, using their semantic category. We sorted the concepts within each group using distance and co-occurrence. We then focussed on the *Physiology* group, using our expert knowledge that most relevant concepts should be in that group.

5.3.2 One-step approach

This process starts with downloading a large, general set of MEDLINE records. This set of documents is converted into a set of fingerprints by Collexis software. Using the resulting set of fingerprints, an ACS is trained. The concepts in this ACS are the candidates for C-concepts in a A to C discovery process. A ranked and filtered list of these candidates is extracted from the ACS. We will evaluate the different rankings by examining the positions in the ranking of the relevant concepts for C. We can not use ROC analysis here, because the number of candidates is too large for us to label as we did with the B-candidates in the two-step approach.

Basic approach

The basic approach starts with downloading MEDLINE records using the following query:

```
(raynaud s disease OR diet OR dietary) AND 1900:1985[dp]
```

Table 5.5: ACS parameter values for the one-step approach

number of dimensions	n	8
constant for learning rate	u	10
number of learning cycles	T	150
fingerprint cutoff	$w_{\min}$	0.5

The settings used to train an ACS on this document set are listed in table 5.5. The number of learning cycles is much higher because of the larger size of the fingerprint set. The cutoff is higher to keep the size of the ACS within bounds. A higher cutoff results in fewer concepts and fewer edges. All concepts in this ACS are regarded C-candidates. We sorted them by distance to A.

Using inverse document frequency

We use the same fingerprint collection as in the basic one-step approach and the same method of correcting for the IDF as used with the two-step approach.

Using semantic categories

We applied the same semantic groups used in the two-step approach to both the ACS used in the basic one-step approach and the ACS used in the IDF-corrected one-step approach. The group of interest is now *Chemicals & Drugs*.

Chapter 6

Test results

The results of various experiments done using the test case and evaluation method described in the previous chapter are described here. The first section discusses the results of the experiments done with the two-step approach, which were described in section 5.3.1, the second section discusses those done with the one-step approach, which were described in section 5.3.2.

6.1 Two-step approach

During experimentation with the two-step approach, we observed that the top 10 highly relevant concepts usually contain concepts from all three path ways. We think that when the user has studied multiple concept from more than one of these pathways, a discovery of the $A - B$ step is likely. We therefore assume that when the user has seen roughly one fourth of the highly relevant B-concepts (10 out of 39), discovery is likely. The other highly relevant B-concepts can be identified later using a closed discovery process. We will present the 10 highest ranking concepts with their positions for each experiment. The names of the concepts are presented for illustration. The position tables also contain the results of ranking by co-occurrence.

To compare the various modifications with the basic approach, we present AUC tables with each experiment. These contain the AUC of the ROC curves based on highly relevant concepts and those based on all relevant concepts (see section 5.1.2). They also contain AUC values based on smaller values of n (see section 5.2). With some experiments, we will show the ROC curves for illustration. The AUC tables and ROC curves will also contain the results of ranking by co-occurrence, with which we will evaluate the contribution of the ACS.

Table 6.1: Highest ranking highly relevant concepts for experiment 1-1

Distance		Co-occurrence	
Position	Name	Position	Name
1	Vasodilation	53	Blood Pressure
93	Blood Circulation	54	Blood Flow Velocity
104	Anticoagulants	70	Vasoconstriction
115	Thrombosis	74	Thrombosis
209	Blood Flow Velocity	84	Vasodilation
253	Dipyridamole	91	Blood Viscosity
266	Blood Pressure	96	Blood Circulation
289	Blood Viscosity	153	Epoprostenol
293	Vascular Resistance	208	Vascular Resistance
320	Capillary Resistance	231	Platelet Aggregation

For each experiment, we will assess the chance a relevant concept has to be ranked higher than a non-relevant one. This chance is estimated using both the AUC obtained using highly relevant concepts and the AUC obtained using all relevant concepts.

6.1.1 Experiment 1-1: The basic approach

The document set used in this experiment contains 2,176 documents. The ACS trained on this set contained 2,336 concepts, with a total of 49,727 edges between them. Of the 2,335 B-candidates from this ACS, 39 are highly relevant. The highest ranking highly relevant concepts are listed in table 6.1 along with their positions in the ranking. The positions in the distance ranking vary between 1 and 320. This means that the user will have to study 320 concepts to discover one fourth of the highly relevant concepts.

The ROC curves based on highly relevant concepts are plotted in figure 6.1 and those based on all relevant concepts in figure 6.2. The different values for the AUC of these curves are shown in table 6.2.

The AUC of both distance curves vary roughly between 0.64 and 0.69. This means that a relevant concept has a chance of being ranked higher than a non-relevant concept roughly between 64% and 69%. Rankings produced by this basic approach thus do better than random (a random ranking would have a AUC of 50%). Comparing distance and co-occurrence, we observe that both for the results based on highly relevant concepts and the results

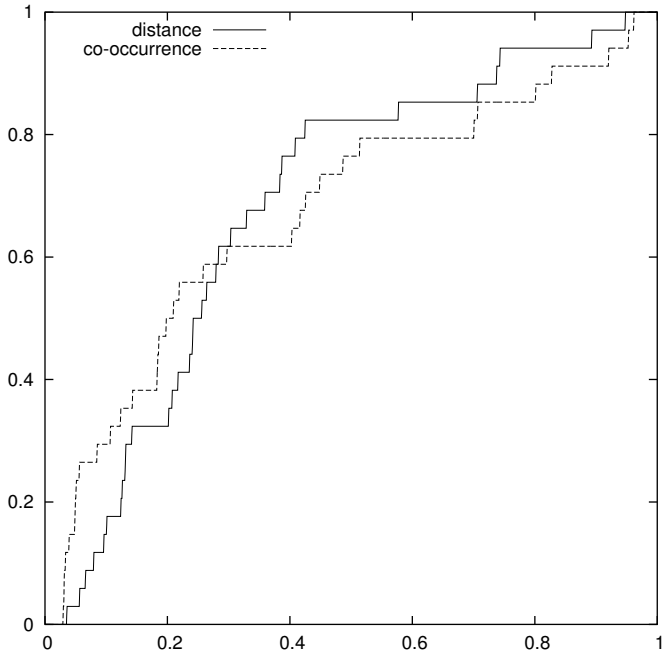

Figure 6.1: Exp. 1-1 ROC: Based on highly relevant concepts.

based on all relevant concepts, distance has a larger AUC. Co-occurrence has higher AUC values in the first part of the curve, for which distance compensates in the rest of the curve. This is especially noticeable when looking at figure 6.1.

6.1.2 Experiment 1-2: Combining rankings

The rankings of B-candidates based on average candidate weight W or average seed weight S, have lower AUC values than those based on distance or co-occurrence. This is shown in figure 6.3. However, several combinations of rankings have a similar or better AUC than any of the single rankings. An interesting combination is that of D and C, which does better than either distance or co-occurrence alone. Another combination that has high AUC values is that of D, C, and W.

Both these combinations are plotted in figure 6.4. Their AUC values are listed in table 6.3. The AUC values of all the combinations tested can be found in the appendix. Table 6.4 lists the top ranking highly relevant

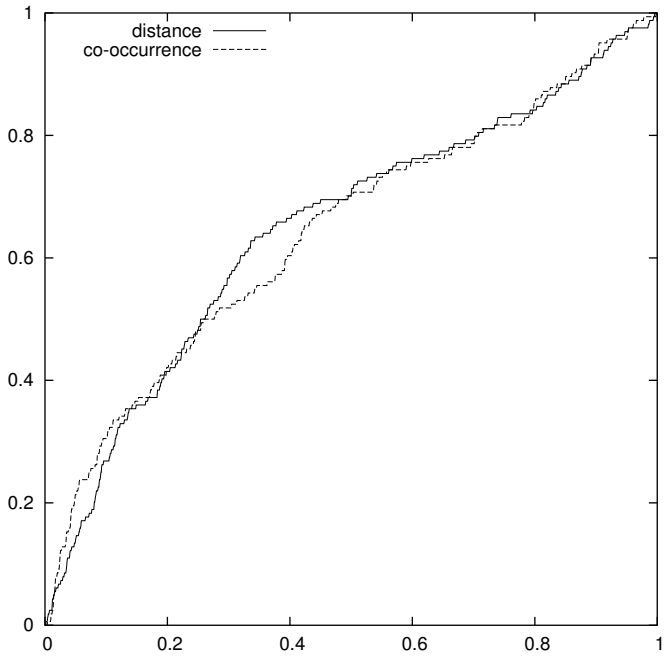

Figure 6.2: Exp. 1-1 ROC: Based on all relevant concepts.

Table 6.2: AUC for experiment 1-1

	Ranked on	Auc	1/5	2/5	3/5	4/5
Highly relevant	Distance	0.685	0.033	0.151	0.315	0.492
	Co-occurrence	0.673	0.055	0.172	0.322	0.487
All relevant	Distance	0.647	0.048	0.160	0.303	0.463
	Co-occurrence	0.641	0.054	0.157	0.297	0.455

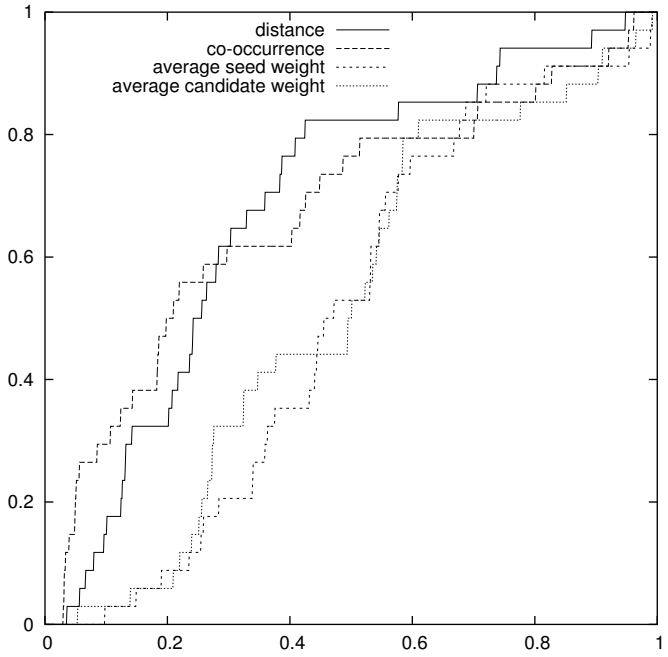

Figure 6.3: Exp. 1-2 ROC 1

concepts in these two rankings.

6.1.3 Experiment 1-3: Using a more general document set

Here, we use a more general document set to see if the ACS-distance will improve. The document set used contained 23,749 documents. The ACS trained on this set contained 6,531 concepts with 313,818 edges. The AUC values of the rankings based on distance and co-occurrence are listed in table 6.5.

Note that we do not present the highest ranking concepts for this experiment. This because it is based on a different document set, and the resulting positions of relevant concepts are not comparable with the positions in other experiments.

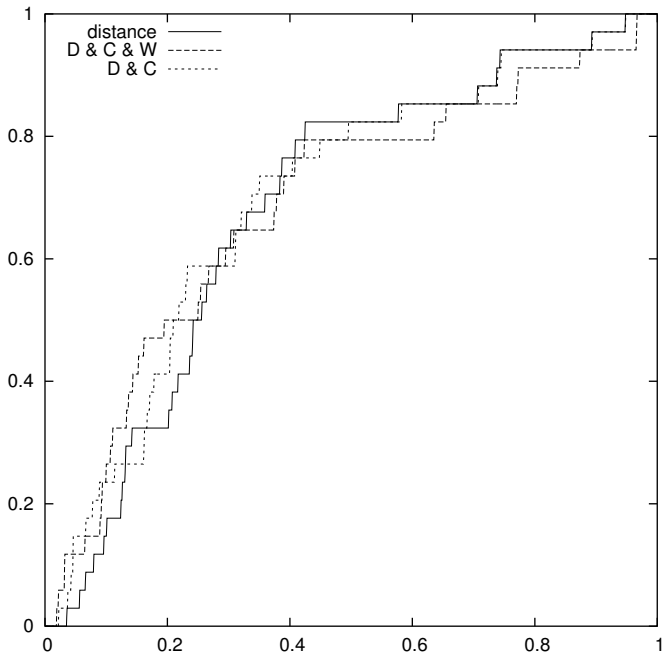

Figure 6.4: Exp. 1-2 ROC 2

Table 6.3: AUC for experiment 1-2

	Ranked on	Auc	1/5	2/5	3/5	4/5
Highly relevant	D	0.685	0.032	0.146	0.311	0.490
	DC	0.698	0.042	0.168	0.329	0.505
	DCW	0.686	0.051	0.171	0.329	0.498
All relevant	D	0.647	0.042	0.146	0.293	0.459
	DC	0.652	0.054	0.163	0.304	0.466
	DCW	0.658	0.061	0.172	0.311	0.472

Table 6.4: Highest ranking highly relevant concepts for experiment 1-2

D and C		D, C, and W	
Position	Name	Position	Name
6	Vasodilation	15	Vasodilation
32	Thrombosis	31	Thrombosis
35	Blood Circulation	38	Blood Circulation
61	Blood Flow Velocity	49	Blood Flow Velocity
88	Blood Pressure	62	Blood Pressure
107	Blood Viscosity	87	Blood Viscosity
152	Vasoconstriction	112	Vasoconstriction
180	Vascular Resistance	179	Vascular Resistance
223	Platelet Aggregation	198	Platelet Aggregation
282	Anticoagulants	259	Epoprostenol

Table 6.5: AUC for experiment 1-3

	Ranked on	Auc	1/5	2/5	3/5	4/5
Highly relevant	Distance	0.630	0.023	0.108	0.255	0.439
	Co-occurrence	0.596	0.044	0.137	0.247	0.403
All relevant	Distance	0.628	0.028	0.115	0.258	0.437
	Co-occurrence	0.669	0.058	0.170	0.309	0.476

Table 6.6: AUC for experiment 1-4

	Ranked on	Auc	1/5	2/5	3/5	4/5
Highly relevant	Distance	0.702	0.041	0.157	0.315	0.502
	Co-occurrence	0.658	0.048	0.158	0.301	0.469
All relevant	Distance	0.730	0.061	0.202	0.360	0.537
	Co-occurrence	0.677	0.062	0.171	0.318	0.488

Table 6.7: Highest ranking highly relevant concepts for experiment 1-4

Distance		Co-occurrence	
Position	Name	Position	Name
36	Thrombosis	42	Blood Viscosity
80	Platelet Aggregation	55	Vasoconstriction
82	Vasodilation	60	Blood Circulation
110	Viscosity	64	Epoprostenol
113	Blood Coagulation	65	Blood Flow Velocity
169	Platelet Adhesiveness	70	Viscosity
215	Alprostadil	93	Vasodilation
263	Blood Pressure	105	Thrombosis
287	Fibrinolysis	148	Platelet Aggregation
289	Blood Viscosity	190	Alprostadil

6.1.4 Experiment 1-4: Using inverse document frequency

The ACS based on the corrected fingerprint set has 2,223 concepts, with 26,439 edges. The AUC values for distance and co-occurrence are listed in table 6.6.

The AUC values of both distance and co-occurrence have increased when compared with experiment 1-1. However, the AUC of the distance ranking has increased relatively more, and distance has higher AUC values than co-occurrence in this experiment. The top ranking concepts are listed in table 6.7.

6.1.5 Experiment 1-5: Using semantic categories

Figures 6.5 and 6.6 and table 6.8 show the ROC curves and AUC values of distance and co-occurrence within the 'Physiology' group. The top ranking concepts within this category are listed in table 6.7. The group contains

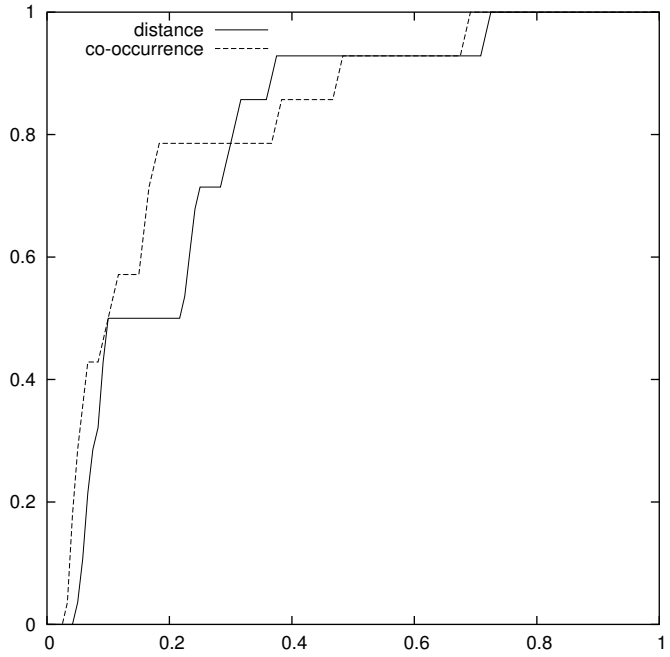

Figure 6.5: Exp. 1-5 ROC: Using highly relevant concepts

a total of 143 candidate concepts. Both the AUC values and the positions of the top ranking concepts are much higher than in any of the preceding experiments.

6.2 One-step approach

For the one-step approach, we will study the rankings of the C-candidates by the basic approach and its modifications. We will show the positions of the relevant C concepts. Of these, the concepts 'Dietary Fats' and 'Fatty Acids, Essential' are more general than the other five. They are relevant, but we think that about three of the other five should be seen by the user to make discovery likely.

6.2.1 Experiment 2-1: The basic approach

The document set used in this basic one-step approach contains 105,928 documents. The ACS resulting from this set has 11,072 concepts with 707,634

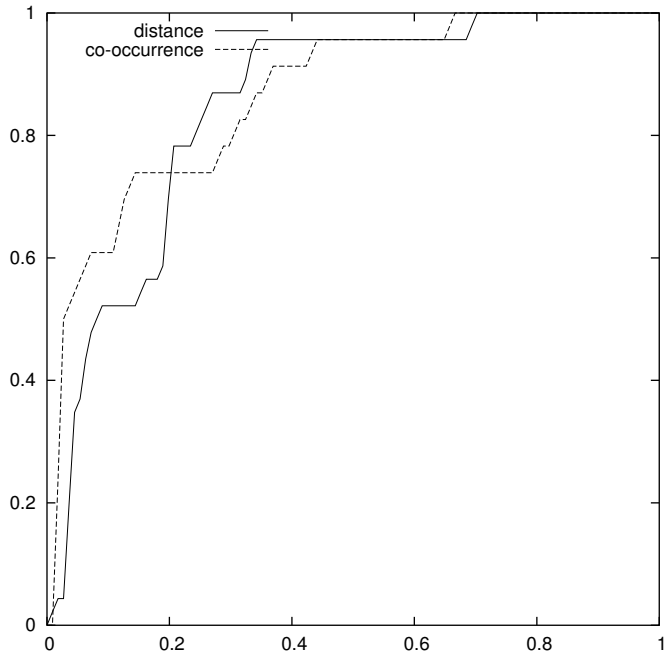

Figure 6.6: Exp. 1-5 ROC: Using all relevant concepts

Table 6.8: AUC for experiment 1-5

	Ranked on	Auc	1/5	2/5	3/5	4/5
Highly relevant	Distance	0.793	0.068	0.224	0.410	0.602
	Co-occurrence	0.821	0.095	0.254	0.435	0.630
All relevant	Distance	0.847	0.090	0.265	0.455	0.649
	Co-occurrence	0.865	0.123	0.284	0.472	0.668

Table 6.9: Highest ranking highly relevant concepts for experiment 1-5

Distance		Co-occurrence	
Position	Name	Position	Name
6	Blood Circulation	4	Blood Flow Velocity
8	Vasodilation	6	Blood Pressure
10	Blood Flow Velocity	7	Blood Viscosity
11	Blood Pressure	8	Vasoconstriction
14	Erythrocyte Deformability	11	Blood Circulation
16	Platelet Aggregation	12	Vasodilation
17	Vasoconstriction	17	Platelet Aggregation
34	Fibrinolysis	20	Fibrinolysis
36	Platelet Adhesiveness	27	Platelet Adhesiveness
38	Vascular Resistance	28	Vascular Resistance

Table 6.10: Relevant concepts in experiment 2-1

Position	Name
55	Dietary Fats
301	Fatty Acids, Essential
1278	Fish Oils
2133	Cod Liver Oil
2202	Docosahexaenoic Acids
3721	Fatty Acids, Omega-3
4438	5,8,11,14,17-Eicosapentaenoic Acid

edges. The positions in the distance ranking of the 7 relevant C-concepts in this ACS are listed in table 6.10.

6.2.2 Experiment 2-2: Using inverse document frequency

In this experiment, the fingerprints are corrected with IDF. The ACS based on the corrected fingerprints contains 11,984 concepts and 564,125 edges. The positions of the relevant concepts in the distance ranking are listed in table 6.11.

Table 6.11: Relevant concepts in experiment 2-2

Position	Name
535	Dietary Fats
574	Fatty Acids, Essential
1188	Fish Oils
2438	Cod Liver Oil
3562	Docosahexaenoic Acids
3834	Fatty Acids, Omega-3
4803	5,8,11,14,17-Eicosapentaenoic Acid

Table 6.12: Relevant concepts in experiment 2-3-1

Position	Name
119	Fatty Acids, Essential
487	Fish Oils
781	Cod Liver Oil
805	Docosahexaenoic Acids
1340	Fatty Acids, Omega-3
1612	5,8,11,14,17-Eicosapentaenoic Acid

6.2.3 Experiment 2-3: Using semantic categories

We applied semantic categories to both the ranking produced by the basic approach and the ranking produced by the IDF-corrected approach. We focus on the category 'Chemicals & Drugs', which contains 6 out of the 7 relevant concepts.

In the ranking based on the basic approach, Chemicals & Drugs contains 3,576 concepts. The positions of the C-concepts in this category are listed in table 6.12.

In the ranking based on the IDF-corrected approach, Chemicals & Drugs contains 4,240 concepts. The positions the C-concepts in this category are listed in table 6.13.

Table 6.13: Relevant concepts in experiment 2-3-2

Position	Name
182	Fish Oils
183	Fatty Acids, Essential
1199	Cod Liver Oil
1284	Docosahexaenoic Acids
1304	Fatty Acids, Omega-3
1647	5,8,11,14,17-Eicosapentaenoic Acid

6.3 Summary of results

Two-step approach

In the ranking produced by our basic two-step approach (section 6.1.1), the 320 highest ranking B-candidates contain one fourth of the highly relevant B-concepts. The chance that a relevant concept is ranked higher than a not relevant one is between 64% and 69%. The ranking based on distance has a higher AUC than the ranking based on co-occurrence, but the difference is very small. The co-occurrence ranking does better for smaller values of n.

The best combined rankings (section 6.1.2) are DC and DCW. For DC, the chance of highly relevant concepts being ranked higher than other concepts is between 65% and 70%, for DCW this chance is between 66% and 67%. The user will have to study 282 (DC) and 259 (DC) concepts to discover one fourth of the highly relevant concepts. The AUC values of these rankings are higher than the AUC values of distance in the basic approach (D). Interestingly, the AUC of DC is higher than the AUC of both D and C. For smaller values of n, DC and DCW also do better than D, if only marginally so.

In the third experiment (section 6.1.3), we tried a more general document set. In the ranking based on this, the chance of relevant concepts being ranked higher than other concepts is roughly 63%. The AUC values of the rankings based on highly relevant concepts suggest that distance does better than co-occurrence in a larger ACS. When we include all relevant concepts however, the roles are reversed. In both cases, the values are lower than those obtained with the basic approach.

The results of using inverse document frequency (section 6.1.4) are better, with a chance of 70% to 73% of highly relevant concepts being ranked higher than other concepts. The user will still have to study 289 concepts

to discover the 10 highest ranked highly relevant concepts. The AUC values of this approach are higher than those obtained with the basic approach. Using co-occurrence, we also obtain higher rankings. However, the AUC of the distance ranking has increased relatively more, and distance has higher AUC values than co-occurrence in this experiment.

In section 6.1.5, we use semantic categories to focus on the semantic category 'Physiology'. With this focus, the number of candidates is reduced to 143 concepts, of which the user has to study 38 to discover the 10 highest ranking highly relevant concepts. The chance of a relevant concept ranked higher than a not relevant concept in this group is between 79% and 85%. The AUC values within this category are much higher than in the basic approach. The values of the co-occurrence ranking have increased even more and are higher than distance in this experiment.

One-step approach

In the ranking produced by the basic one-step approach (6.2.1), the user will have to study 2,202 out of 11,072 concepts to discover three of the five more specific relevant concepts.

When using IDF-corrected fingerprints (6.2.1), the user will have to study 3,562 concepts for this.

When we focus on the semantic category 'Chemicals & Drugs', the number of concepts the user has to study is 805 out of 4,240 using the basic ranking and 1,284 using the IDF-corrected ranking.

Chapter 7

Discussion and outlook

In this chapter we discuss our results and methodology, and we look to the future. In section 7.1, the results presented in chapter 6 are discussed. The limitations of our evaluation are discussed in section 7.2. Next, we do some suggestions for further research in section 7.3. In section 7.4, we present an out look to the future of literature-based knowledge discovery. Finally, section 7.5 contains some acknowledgements.

7.1 Discussion of results

Chapter 5 identified three objectives of the evaluation of these suggestions:

1. To evaluate the added value of the ACS in the discovery process.
2. To evaluate whether any of the possible modifications improve the basic discovery process and thus should be used.
3. To evaluate whether a scientist using either the two-step or the one-step approach will be able to make discoveries with a reasonable amount of effort.

Here, we discuss the results presented in chapter 6 to meet these three objectives.

Contribution of the ACS

We evaluated the added value of the ACS in the discovery process by comparing the rankings obtained by using ACS-distance with rankings obtained using co-occurrence. Since the C-candidates do not co-occur with A, this can

only be done for the two-step approach and not for the one-step approach. We compared the rankings by their AUC-values.

For the basic two-step approach and for the modification which uses a more general document set, we have contradicting results. Some results indicate the ranking based on distance is better, some that the ranking based on co-occurrence is better. In both cases, the differences are very small. The distance ranking improves more from IDF-corrected fingerprints than co-occurrence. The results using these fingerprints indicate that distance does better than co-occurrence. When we focus on the semantic category 'Physiology', co-occurrence has better results than distance.

In two experiments, the difference between co-occurrence and distance is small or not clear. On one, distance does better. In another, co-occurrence does better. Based on these four experiments, we conclude that the ACS has no added value in a two-step discovery process.

Using modifications

To evaluate whether any of the possible modifications improve the basic discovery process and thus should be used, we compared the rankings obtained by the basic approach with those produced by the modifications. For the two-step approach, we looked at the AUC values. We also looked at the number of concepts that the user has to study to make discovery likely. For the one-step approach, we looked at the positions of the relevant concepts.

In the first modification done on the two-step approach, there are some combinations of features which have better results than the distance used in the basic approach. However, the differences are small. The results of using a more general document set are worse than those of the basic approach. When we use IDF-corrected fingerprints, the results are better than those of the basic approach. Focussing on a semantic category improves the results drastically.

For the one-step approach, we tried two modifications. The results obtained with IDF-correction are worse than those obtained with the basic approach. The positions obtained with the use of semantic categories are higher. However, this is solely caused by a decrease in candidate concepts. The user still has to study roughly one fifth of the candidates to make the discovery.

We conclude that our first two modifications to the two-step approach do not improve the discovery process. The last two do. This suggests that IDF-correction should be used for the two-step approach. Using semantic categories drastically improves results. They should be used when the

user has the required expert knowledge to focus on one or more semantic categories.

In our experiments with the one-step approach, IDF correction does not improve the discovery process. Using semantic categories succeeds in filtering the candidates, but the ranking within the semantic category does not improve. Again, they should be used when the user has the required expert knowledge.

Making discoveries

To see if literature-based knowledge discovery is possible with one of our approaches, we looked at the number of concepts that have to be studied to make discovery likely. For the two-step approach, we also looked at the chance a relevant concept has to be ranked higher than a not relevant one. For both approaches, we discuss both the best method (basic or modified) which does not require the expert knowledge to focus on a semantic category and the method with semantic categories.

In the IDF-corrected two-step approach, the user will have to study almost 300 concepts to make discovery likely. The chance of a relevant concept being ranked higher than a not relevant one is 70% to 73%. With semantic categories, this chance is 79% to 85%. The user has to study almost 40 concepts to make discovery likely.

For the basic one step approach, the user has to study over 2000 concepts to make discovery likely. Using semantic categories, this number is still over 800.

We think that the use of a literature-based discovery system is very limited when the user has to study a large amount of concepts to make a discovery. This is the case with the IDF-corrected two-step approach and the one-step approach. The two-step approach with semantic categories seems suitable for literature-based discovery.

Summary

We conclude that:

- Our results do not indicate that the ACS contributes to a literature-based discovery process.
- Using inverse document frequency and semantic categories contributes to our two-step discovery process. Semantic categories also contribute to our one-step process.

- Literature-based knowledge discovery is likely possible with our two-step approach with semantic categories without requiring more than a reasonable amount of effort.

7.2 Limitations of our evaluation method

There are several limitations to the way we have evaluated our approaches. First of these is the use of just one study case. It introduces the chance that the conclusions we have drawn from the test results are only valid for this single case, and will not hold when applied to other cases. Possible solutions are the use of artificial data and increasing the number of test cases. It is difficult, however, to capture the complexities of literature in artificial data. More test cases are available, and using them would have improved the confidence in our test results. Time constraints were the reason we did not use them.

The labelling of concepts as relevant or not relevant is another limitation of our method. It is subjective and a different set of relevant concepts may lead to different results and thus different conclusions. We tried to take part of this effect away by defining two different sets of relevant concepts (highly relevant and all relevant concepts), and comparing the results using either one. In one case, this led to contradicting results. In all other cases, the results had the same implications.

Another limitation is the experimental status of the ACS algorithm. The settings we use for ACS training are based upon (**?**) and on some experimentation done by ourselves. However, neither of these sources are out of the experimental phase when the optimal ACS training parameters are concerned. It is not certain that the parameters we used are optimal for ACS training. Thus, it is possible that the ACS-distance rankings are sub-optimal.

We did not use ROC analysis for the one-step approach. This makes it hard to compare the basic approach with its modifications. The reason for this was the limited amount of relevant concepts we defined. We could have solved this by defining more loosely relevant C-concepts and dividing the set as done with the two-step approach. However, the set of candidate C-concepts is very large and time constraints withheld us from spending the necessary time to manually label such a large set of concepts.

The two-step approach was only evaluated based on the $A - B$ step. A proper case evaluation should also do the $B - C$ step. However, we tried the method with which the $A - B$ step was successfully made, semantic

categories, near the end of our research period. Time constraints withheld us from evaluating the $B - C$ step.

7.3 Suggestions for further research

Two-step approach with semantic categories

We obtained the best results while using semantic categories in a two-step discovery process. We suggest further examination of the possibilities of this approach. Follow-up work could implement the two-step approach with a number of differences with the basic approach we presented. It should use IDF-corrected fingerprints, and use co-occurrence and semantic categories to rank candidate concepts. The possibilities of this approach can then be further explored by replicating more of Swanson's discoveries.

Note that the system suggested above is very similar to Srinivasan's system for literature-based discovery (section 3.2.5). The differences are the use of free text and fingerprints instead of MeSH headings, and the use of aggregated semantic categories instead of the much narrower semantic types.

One-step approach with semantic categories

We did not reach satisfying results with the one-step approach. Using semantic categories improved the results, but did not lead to a ranking in which discovery is likely. A reason for this might be that the semantic category used there, Chemicals & Drugs, contains many concepts.

We suggest exploring the use of our one-step approach with semantic categories. The categories should be much smaller that the categories we used. In that way, the user can provide (and thus also needs to have) more expert knowledge to the system, and discovery may become likely.

Other suggestions

We tried to exploit the information contained in the fingerprint weights in one of our experiments. We suggest trying different ways to exploit this information. Our last suggestion is to pay special attention to the size and generality of the document sets in which discoveries are done. This has a large influence on the discovery process.

7.4 Outlook

The results obtained with semantic categories suggest that using expert knowledge contributes significantly to the discovery process. Most of the literature-based knowledge discovery systems discussed in chapter 3 also use expert knowledge in their systems. Swanson's system relies on the user to filter terms in the discovery process. In Weeber's system, the user has to select semantic categories to filter concepts. Hristovski also supports this semantic filtering. Also, his system uses human assigned MeSH terms to assess the contents of documents. The best results are obtained by Srinivasan, whose system also relies on human assigned MeSH terms, and does filtering by having the user select semantic types.

We think human expert knowledge is a necessary contribution to the discovery process. There are two reasons for this. In the first place, it is extremely difficult to model all the domain knowledge humans use to contribute to the discovery process. Our results support this. We were only able to successfully identify the relevant B-concepts with the help of expert knowledge of which semantic category to focus on. Even when we could model this knowledge, a fully automated process would not be useful. Such a process would only transform the large amounts of text into large amounts of (ranked) hypotheses. A human is still needed to decide which hypotheses to text experimentally. (See also (Weeber, 2003))

We think therefore that further research in literature-based knowledge discovery should focus on its role as a support system. Combining expert knowledge from the user with computational power from the computer should be the prime target of future systems. We think that in this role, literature-based knowledge discovery has rising prospects.

There are several developments that contribute to these rising prospects. The first is the availability of the scientific literature, which should only increase. Another is increasing computer power, which will enable computers to do more complex analysing of larger amounts of text. We believe that these developments will eventually make literature-based discover a powerful way to assist scientists with hypothesis generation

To have them actually be used (outside of information science), however, the ideas of literature-based discovery will have to be communicated to other fields than that of information science. In (Spasser, 1997), Mark A. Spasser explores the degree to which Swanson's valid idea of using the scientific literature to generate new knowledge has been adopted in the biomedical literature. He concluded that biomedical scientists largely ignored Swanson's ideas. Spasser identified several barriers which may have caused this failure

of exporting Swanson's ideas to the biomedical field.

One of these barriers is identical to the reason for the existence of undiscovered public knowledge: scientist's limited fields of interest. Understanding the possibilities of Swanson's ideas requires at least a basic understanding of information science. However, the scientists in the biomedical field are already busy enough with the knowledge in their own field of interest.

Another barrier he identified consists of the different disciplinary conceptions of what constitutes valid data generation, evaluation, and experimentation. Biomedical researchers (among others) require all theories to be testable in the real world. They are reluctant "to concede the legitimacy of an explicitly exploratory methodology, one that does not depend on empirical and quantitative hypothesis testing".

To overcome these obstacles, it is even more important that literature-based discovery focusses on its role as a support system. Systems should be supported by an extensive user interface and support from information scientists. This takes away the need for users to study the underlying theories. Their skepticism about the possibilities can be taken away by presenting a system that simply works. An example of this is the access to the electronical biomedical library MEDLINE through PubMed. The algorithms with which this database is searched, may be less complex than those used in literature-based knowledge discovery, but understanding them does require knowledge in information science. Knowledge which most users, from the biomedical field, do not have. Yet, they use the system because it is easy to use and because it works.

As technical developments and research make literature-based knowledge discovery more powerful, the likelihood of a user-friendly, working system that assists scientist with hypothesis generation increase. Literature-based knowledge discovery can not replace, but it can support hypothesis driven experimental research. Tools such as we strived to develop here could help scientists develop hypotheses more efficiently and thus improve scientific progress.

7.5 Acknowledgements

I have worked long on this thesis, and although I must admit there were many times when I did not work very hard, there were also lots of times when I did.

I am, however, not the only one who has spent time and effort on this thesis. I would like to thank my supervisors Marc Weeber, Jan Kors, and

Jan van den Berg. Al three of them, and especially Marc, have spent a lot of time advising me, discussing with me, and reading and re-reading my texts.

I would also like to thank some of my co-workers at the Erasmus M.C., where I wrote this thesis. Both Rob Jelier and Christiaan van der Eijk have provided quick and easy answers to several questions I would have had to spend much more time on otherwise.

Finally, I would like to thank some people who have provided more indirect contributions. My parents have been a source of support in many ways, as has my girlfriend Jorine. Although she distracted me from my work on numerous occasions, she had a large part in keeping my spirits high.

Bas van der Lans
March 18, 2004

Bibliography

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of the 2001 AMIA annual fall symposium*, 17–21. Philadelphia, PA: Hanley and Belfus.

Biosemantics Group Rotterdam (2003). Website. http://www.biosemantics.com/.

Chang, B. B., DiGiacomo, R. A., Kremer, J. M., Kay, C., & Shah, D. M. (1988). Effects of fish oil fatty acid ingestion in patients with Raynaud's syndrome. *Surgical Forum*, *39*, 312–318.

DiGiacomo, R. A., Kremer, J. M., & Shah, D. M. (1989). Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *American Journal of Medicine*, *86*, 158–164.

Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for data mining researchers* (Technical Report). Hewlett-Packard Development Company, L.P.

Gordon, M. D., & Lindsay, R. K. (1996). Toward 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*, 116–128.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2003). Improving literature based discovery support by genetic knowledge integration. *Stud Health Technol Inform*, *95*, 68–73.

Hristovski, D., Stare, J., Peterlin, B., & Dzeroski, S. (2001). Supporting discovery in medicine by associating rule mining in Medline and UMLS. *Proceedings of MedInfo*, *10*, 1344–1348.

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

McCray, A. T., Burgun, A., & Bodenreider, O. (2001). Aggregating umls semantic types for reducing conceptual complexity. *Medinfo*, *10(Pt 1)*, 216–220.

Merrian-Webster (2004). Website. http://www.m-w.com/.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

Ming, L. (2002). *Brief report: ROC analysis in machine learning* (Technical Report). University of Bristol, Dept of Computer Science.

National Library of Medicine (2003a). Linkout journals by titles. http://www.nlm.nih.gov/.

National Library of Medicine (2003b). Website. http://www.nlm.nih.gov/.

Schuemie, M. J. (1998). Associatieve conceptuele ruimte, een vorm van kennisrepresentatie ten behoeve van informatie-zoeksystemen. Master's thesis, Erasmus University of Rotterdam.

Schuemie, M. J., & van den Berg, J. (1998). *Associative conceptual space-based information retrieval systems* (Technical Report). Erasmus University of Rotterdam.

Schuemie, M. J., & van den Berg, J. (1999). Information retrieval systems using an associative conceptual space and self-organising maps. *Proceedings of the BNAIC'99* (pp. 91–98).

Smalheiser, N. R., & Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neuroscience Research Communications*, *15*, 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*, 809–810.

Smalheiser, N. R., & Swanson, D. R. (1998a). Calcium-independent phospholipase A2 and schizophrenia. *Archives of General Psychiatry*, *55*, 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.

Spasser, M. A. (1997). The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science*, *48*, 707–717.

Srinivasan, P. (2001). MeSHmap: A text mining tool for MEDLINE. *Proceedings of the American Medical Informatics Annual Symposium* (pp. 642–646).

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *50*, 396–413.

Srinivasan, P., & Sehgal, A. K. (2003). Mining MEDLINE for similar genes and similar drugs. Unpublished.

Srinivasan, P., & Wedemeyer, M. (2003). Mining concept profiles with the vector model or Where on earth are diseases being studied? *Proceedings of the Text Mining Workshop. Third SIAM International Conference on Data Mining.*.

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

Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, *38*, 228–233.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, *31*, 526–557.

Swanson, D. R. (1989). A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science*, *40*, 432–435.

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

Swanson, D. R. (1991). Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiaramella, G. Salton and V. V. Raghavan (Eds.), *Proceedings of the 14th annual international ACM/SIGIR conference on research and development in information retrieval*, 280–289. New York: ACM Press.

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. (1999). Link analysis of MEDLINE titles as an aid to scientific discovery: Using arrowsmith as an aid to scientific discovery. *Library Trends*, *48*, 48–59.

van den Berg, J., & Schuemie, M. J. (1999). Information retrieval systems using an associative conceptual space. *Proceedings of the ESANN'99* (pp. 351–356).

van der Eijk, C. C. (2001). Knowledge discovery in scientific literature. Master's thesis, Erasmus University of Rotterdam.

van der Eijk, C. C., van Mulligen, E. M., 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*, *50*, 436–444.

van der Eijk, C. C., van Mulligen, E. M., & van den Berg, J. (2002). Finding complementary scientific concepts using a conceptual associative spatial graph. *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002).*

van Mulligen, E. M., Diwersy, M., Schmidt, M., Buurman, H., & Mons, B. (2000). Facilitating networks of information. In *Proceedings of the 2000 AMIA annual fall symposium*, 868–872.

van Mulligen, E. M., van der Eijk, C., Kors, J. A., Schijvenaars, B. J. A., & Mons, B. (2002). Research for research: Tools for knowledge discovery and visualization. In *Proceedings of the 2002 AMIA annual fall symposium*, 835–839.

Weeber, M. (2003). Drug discovery as an example of literature-based discovery. Book chapter for *Computational discovery* by Dzeroski and Todorovski.

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 *Proceedings of the 2000 AMIA annual fall symposium*, 903–907. Philadelphia, PA: Hanley and Belfus.

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*, 548–557.

Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T. W., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature. a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association*, *10*, 254–262.

Wyatt, J. (1991). Use and sources of medical knowledge. *338*, 1368–1373.

Appendix

Table A.1: Relevant B-concepts, excluding highly relevant concepts.

Blood Viscocity	
Cryoglobulinemia	Cryoglobulins
Erythrocyte Membrane	Erythromelalgia
Hemangioendothelioma	Hematologic Diseases
Hemodynamics	Hemolysins
Hemophilia A	Hemorheology
Hemorrhagic Disorders	Hemostasis
Hemostatics	Pentoxifylline
Rheology	
Platelet Aggregation	
Agglutination Tests	Albumins
Anemia	Antifibrinolytic Agents
Antithrombin III	Arachidonic Acid
Arachidonic Acids	Blood Cell Count
Dextrans	Dicumarol
Disseminated Intravascular Coagulation	Embolism
Factor VIII	Factor XII
Factor XIII	Fibrin Fibrinogen Degradation Products
Fibrinogen	Gabexate
Heparin	Kallikreins
Ketanserin	Pentoxifylline
Platelet Count	Platelet Function Tests
Prothrombin Time	Thrombocytopenia
Thrombocytosis	Thromboembolism
Thrombophlebitis	Thromboplastin
Thromboxane A2	von Willebrand Factor
Vascular Reactivity	
Aminorex	Angiotensin II

Angiotensin-Converting Enzyme Inhibitors	Antihypertensive Agents
Arterial Occlusive Diseases	Arterioles
Arteriosclerosis	Arteriosclerosis Obliterans
Arteriovenous Anastomosis	Arteriovenous Fistula
Arteriovenous Malformations	Bencyclane
Bradykinin	Capillaries
Captopril	Dihydroergotamine
Diltiazem	Endothelium
Endothelium, Vascular	Epinephrine
Ergoloid Mesylates	Ergonovine
Ergotamine	Fibromuscular Dysplasia
Guanethidine	Histamine
Hydralazine	Hydroxyethylrutoside
Hypotension	Iloprost
Indapamide	Indoramin
Isosorbide Dinitrate	Isoxsuprine
Ketanserin	Labetalol
Microcirculation	Minoxidil
Moxisylyte	Muscle, Smooth, Vascular
Nafronyl	Neurokinin A
Niacin	Nicergoline
Nicotinyl Alcohol	Nifedipine
Nitroglycerin	Nitroprusside
Norepinephrine	Oxprenolol
Papaverine	Pentoxifylline
Phenoxybenzamine	Phentolamine
Phenylpropanolamine	Pindolol
Polyarteritis Nodosa	Prazosin
Propranolol	Sotalol
Suloctidil	Theophylline
Thromboxane A2	Tolazoline
Varicose Veins	Vascular Diseases
Vasoconstrictor Agents	Vasodilator Agents
Vasomotor System	Vasopressins
Xanthinol Niacinate	

Other	
Acidosis, Lactic	Blood Bactericidal Activity
Blood Component Removal	Body Temperature
Body Temperature Regulation	CREST Syndrome

Extracorporeal Circulation	Extremities
Fingers	Ischemia
Pain Insensitivity, Congenital	Peripheral Vascular Diseases
Piribedil	Plasma
Plasma Cells	Plethysmography
Regional Blood Flow	Skin Diseases, Vascular
Skin Temperature	Toes

Table A.2: AUC for experiment 1-2 (Based on highly relevant concepts)

Ranked on	Auc	1/5	2/5	3/5	4/5
D	0.685	0.033	0.151	0.315	0.492
C	0.673	0.055	0.172	0.322	0.487
W	0.519	0.006	0.064	0.173	0.338
S	0.504	0.005	0.046	0.155	0.321
CW	0.672	0.054	0.172	0.319	0.489
CWS	0.667	0.055	0.172	0.315	0.476
WS	0.517	0.002	0.052	0.163	0.325
$D\&C$	0.698	0.042	0.168	0.329	0.505
$D\&S$	0.626	0.037	0.130	0.259	0.433
$D\&W$	0.652	0.040	0.151	0.299	0.466
$C\&S$	0.640	0.046	0.153	0.292	0.460
$C\&W$	0.639	0.046	0.147	0.284	0.449
$S\&W$	0.519	0.002	0.051	0.164	0.327
$D\&C\&S$	0.673	0.046	0.160	0.307	0.480
$D\&C\&W$	0.686	0.051	0.171	0.329	0.498
$D\&S\&W$	0.599	0.022	0.104	0.241	0.409
$C\&S\&W$	0.587	0.026	0.109	0.236	0.396
$D\&C\&S\&W$	0.648	0.041	0.149	0.290	0.459
$D\&CWS$	0.684	0.043	0.165	0.321	0.494
$D\&CW$	0.683	0.042	0.166	0.326	0.496

Table A.3: AUC for experiment 1-2 (Based on all relevant concepts)

Ranked on	Auc	1/5	2/5	3/5	4/5
D	0.647	0.048	0.160	0.303	0.463
C	0.641	0.054	0.157	0.297	0.455
W	0.555	0.023	0.093	0.208	0.368
S	0.480	0.011	0.046	0.142	0.294
CW	0.655	0.057	0.162	0.304	0.466
CWS	0.639	0.056	0.161	0.291	0.449
WS	0.529	0.006	0.062	0.182	0.340
$D\&C$	0.652	0.054	0.163	0.304	0.466
$D\&S$	0.595	0.035	0.130	0.252	0.409
$D\&W$	0.644	0.054	0.163	0.298	0.458
$C\&S$	0.597	0.040	0.135	0.262	0.415
$C\&W$	0.641	0.055	0.160	0.296	0.455
$S\&W$	0.527	0.006	0.061	0.180	0.338
$D\&C\&S$	0.628	0.049	0.153	0.285	0.441
$D\&C\&W$	0.658	0.061	0.172	0.311	0.472
$D\&S\&W$	0.591	0.029	0.118	0.244	0.402
$C\&S\&W$	0.578	0.025	0.108	0.233	0.388
$D\&C\&S\&W$	0.629	0.052	0.155	0.281	0.440
$D\&CWS$	0.647	0.055	0.163	0.302	0.459
$D\&CW$	0.654	0.055	0.164	0.306	0.468