Text Mining: Generating Hypotheses From MEDLINE

Padmini Srinivasan
School of Library and Information Science, The University of Iowa, Iowa City, IA, 52242.
E-mail: padmini-srinivasan@uiowa.edu

Hypothesis generation, a crucial initial step for making scientific discoveries, relies on prior knowledge, experience, and intuition. Chance connections made between seemingly distinct subareas sometimes turn out to be fruitful. The goal in text mining is to assist in this process by automatically discovering a small set of interesting hypotheses from a suitable text collection. In this report, we present open and closed text mining algorithms that are built within the discovery framework established by Swanson and Smalheiser. Our algorithms represent topics using metadata profiles. When applied to MEDLINE, these are MeSH based profiles. We present experiments that demonstrate the effectiveness of our algorithms. Specifically, our algorithms successfully generate ranked term lists where the key terms representing novel relationships between topics are ranked high.

Introduction

It is well understood that biomedical knowledge is growing at an astounding pace. This creates an enormous challenge for scientists trying to keep pace with developments in their field. At the same time, these vast collections of publications offer an excellent opportunity for text mining, i.e., the automatic discovery of knowledge. Text mining is similar to data mining (Agrawal & Srikant, 1994; Fayyad & Uthurusamy, 1996; Piatetsky-Shapiro & Frawley, 1991) in its goal. But instead of mining a collection of well-structured data, text mining operates off text collections that are at best semi-structured. In both cases, the *knowledge* discovered is essentially a set of propositions or hypotheses that require further study and verification. Text mining has attracted the attention of many researchers (e.g., Andrade & Valencia, 1998; Feldman et al., 1997; Gordon & Lindsay, 1996; Hahn and Schnattinger, 1997; Hearst, 1999; Hristovski et al., 2001; Lent et al., 1997; Masys et al., 2001; Smalheiser & Swanson, 1996a,b; Srinivasan, 2001, Swanson, 1986, 1988; Swanson et al., 2001; Weeber et al., 2000, 2001), including those in biomedicine. A recent article in *Nature* (Blagosklonny & Pardee, 2002) referring to text mining as conceptual biology speaks to its legitimacy as a field that fuels hypothesis-driven biomedical explorations. Examples of recent text-mining applications include automatically identifying viruses that may be used as bioweapons (Swanson et al., 2001), proposing therapeutic uses for thalidomide (Weeber et al., 2003), and finding functional connections between genes (Chaussabel & Sher, 2002; Shatkay et al., 2000).

Our goal in this report is to propose and evaluate text-mining algorithms designed for hypothesis discovery. We follow the discovery framework initiated by Swanson in the mid 1980s. His aim was to process MEDLINE in particular ways and generate interesting hypotheses concerning specific diseases and health problems. Given two topics that are bibliographically disconnected areas of specialization, Swanson explored potential linkages via intermediate topics or specializations. Over the past two decades and in collaboration with Smalheiser, Swanson proposed several interesting hypotheses (Swanson, 1986, 1988; Swanson & Smalheiser, 1997; Swanson et al., 2001; Smalheiser & Swanson, 1996a,b), that were later validated by bioscientists. Since their pioneering contributions, this kind of knowledge discovery work has attracted the attention of other researchers (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999, Weeber et al., 2000, 2001) besides us.

Swanson and Smalheiser's discovery method may be viewed as having two dimensions. Given a particular topic of interest the first is about identifying interesting related concepts and the second is about exploring the particulars of the relationships. Thus, given a disease X as a starting point, the first dimension is about identifying concepts related to X such as a particular drug Y. In the second dimension, the nature of their relationship is explored: is Y likely to treat X or is it likely to aggravate X or does some other kind of relationship potentially hold between them? The emphasis is on novelty, discovering new Y concepts and/or postulating new relationships between X and Y. We refer to these as dimensions since a given knowledge discovery procedure may intertwine them in complex ways.

Received May 22, 2003; revised September 24, 2003; accepted September 24, 2003

Our algorithms concentrate on the first dimension, i.e., on identifying interesting concepts. (We handle the second dimension as do the others, through manual analysis of the literature.) Our motivation is twofold. First, previous efforts within the Swanson and Smalheiser discovery framework primarily use the free-text portions of MEDLINE (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999, Weeber et al., 2000, 2001) with MeSH (Medical Subject Headings)[1], the metadata applied to MEDLINE records, having at best a secondary role. Thus, we would like to determine the effectiveness of a procedure that almost completely relies on MeSH for the first dimension. Our second motivation is one that we share with other researchers in that we would like to reduce the amount of manual effort involved during the discovery process. Specifically, we would like to automatically return ranked lists of concepts to the user with interesting concepts appearing at the top ranks.

Our algorithms operate by building MeSH-based profiles from MEDLINE for topics. A profile is essentially a set of MeSH terms that together represent the corresponding topic. Different kinds of profiles may be constructed based on the kinds of MeSH terms included. Overall, we assume that as far as the first dimension is concerned, the user is only required to specify (1) the initial topic(s) of interest and (2) the kinds of profiles to generate. We, thus, evaluate our algorithms under these conditions. Our long-term goal is to build a suite of text mining tools that may be used by a domain expert to explore a text collection for hypothesis generation.

In this report, we present our text-mining algorithms as well as experiments replicating several of the discoveries made by Swanson and Smalheiser. We also study the effect of varying the few parameters in our algorithms. We organize the paper as follows. Next we present details about our profiles. Following this, we discuss hypothesis generation and present our discovery algorithms. We then present several discovery experiments. After a discussion of results followed by an overview of related research in the next section. We make our conclusions.

Profiles

Consider a topic such as *Marfans syndrome*, which is a hereditary disease. The profile for this topic distilled from a suitable text collection could identify, for example, terms representing the genes, proteins, symptoms, drug treatments, other diseases, and population groups associated with the disease, i.e., "statistically related" to it. Although it is not necessarily true, we assume that a statistical association implies some semantic association. The profile for a topic such as *Jimmy Carter*, extracted from a text collection of Associated Press or Reuters newsfeeds, may include terms representing the nations he visited, the Habitat for Humanity projects he initiated, and the national elections that he has observed.

We build topic profiles by first identifying a relevant subset of documents from the text collection. We then identify characteristic terms (single words and/or phrases) from this subset and assess their relative importance as descriptors of the topic. Terms may be extracted from the free-text portions of the documents or/and from their metadata. In this reports we build profiles from MEDLINE using just the MeSH metadata. MeSH terms are assigned to the records by trained indexers at NLM who select from a MeSH hierarchy of around 21,000 phrases. In essence, our profiles are weighted vectors of MeSH terms as shown below for a topic T_i.

$$Profile(T_i) = \{w_{i,1}m_1, w_{i,2}m_2, \cdots, w_{i,n}m_n\} \quad (1)$$

where m_j represents a MeSH term, $w_{i,j}$ its weight, and there are totally n terms in the MeSH vocabulary. (We discuss weights shortly.)

Profiles may be as current as the text collection. Alternatively, profiles may be generated from collection subsets corresponding to particular time periods. Such temporal profiles may support trend analysis (Feldman & Dagan, 1995; Lent et al., 1997; Srinivasan & Wedemeyer, 2003). Topics profiled may be as simple as those described by single words (e.g., *Tylenol*) or a user may specify more complex topics such as *Calcium channel blockers and Alzheimers disease* or *Climbing expeditions on the K2*. MEDLINE topics may be in free-text format, i.e., not limited to MeSH terms.

Employing Semantic Types in Profiles

Thus far, our profiles are simply vectors of weighted MeSH terms. Now we describe how we are able to further differentiate between the MeSH terms using semantic types. Specifically, we exploit the fact that the MeSH vocabulary has already been classified using 134 UMLS (Unified Medical Language System)[2] semantic types (NLM, 2002). *Cell Function, Sign or Symptom* are two examples of semantic types. Each MeSH term is assigned one or more semantic types. For example, *interferon type II* falls within both *Immunologic Factor* and *Pharmacologic Substance* semantic types. More generally, semantic types represent "categories" that have been used to classify the MeSH metadata. Figure 1 shows a brief example connecting a MEDLINE record, MeSH, and the UMLS semantic types. Henceforth, MeSH terms will be in lowercase while semantic types will have the first character of each word capitalized. Both will appear in italics.

Figure 2, which outlines our procedure for building profiles, shows how we involve these semantic types. Basically, MeSH terms are separated by semantic type and term

[1] http://www.nlm.nih.gov/mesh/meshhome.html

[2] http://umlsks.nlm.nih.gov

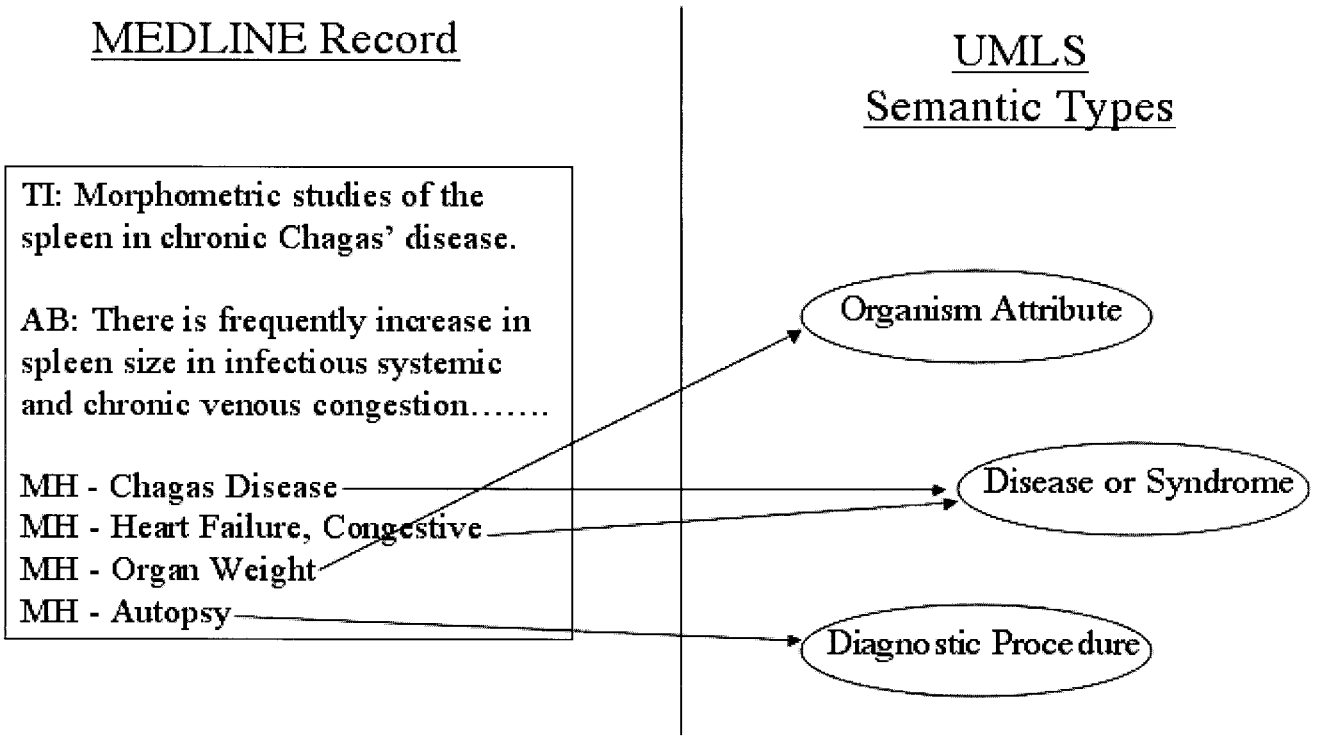

FIG. 1. MEDLINE, MeSH, and UMLS semantic types.

weights are computed within the context of a semantic type. This results in a vector of MeSH term vectors, one for each of the 134 UMLS semantic types. Thus,

$$Profile(T_i) = \{\{w_{i,1,1}m_{1,1}, w_{i,1,2}m_{1,2}, \cdots,\}, \cdots, \{w_{i,134,1}m_{134,1}, w_{i,134,2}m_{134,2}, \cdots,\}\} \quad (2)$$

where $m_{x,y}$ represents the MeSH term m_y that belongs to the semantic type x and $w_{i,x,y}$ is the computed weight for $m_{x,y}$. Weights may be computed using any appropriate weighting scheme (such as mutual information and log likelihood). Below we use the TF*IDF (term frequency * inverse document frequency) (Sparck Jones, 1972) weighting scheme and then normalize the weights:

$$w_{i,x,y} = v_{i,x,y}/highest(v_{i,x,l}), \quad (3)$$

where $l = 1, \ldots, r$ and $v_{i,x,y} = n_{i,x,y} * log(N/n_{x,y})$. Here N is the number of documents in the database, $n_{x,y}$ is the number of documents in which $m_{x,y}$ occurs, and $n_{i,x,y}$ is the number of retrieved documents for T_i in which $m_{x,y}$ occurs. Normalization by *highest* $(v_{i,x,l})$, the highest value for $v_{i,x,y}$ observed for the MeSH terms with semantic type x, yields weights that are in [0,1] within each semantic type. (Note that there are r terms in the domain for semantic type x.)

Thus, a profile represents the relative importance, within semantic types, of the different MeSH terms associated with the topic's document set. When appropriate, this profile may be focussed or limited to a specific *view*, i.e., terms with particular semantic types. For example, profiles of genes

Input from user: a topic X of interest.

1. Conduct an appropriate search for X against MEDLINE and collect the retrieved set R.
2. Extract all MeSH terms from R.
3. For each MeSH term calculate its weight using the TF*IDF scheme. Within each semantic type, normalize the term weights by the highest term weight observed for the semantic type as shown in (equation 3).

Output: Profile for X: A vector of MeSH term vectors one for each semantic type. Terms in each semantic type vector have weights in [0,1].

FIG. 2. Procedure for generating TF*IDF weighted MeSH profiles.

TABLE 1. Example profile. Topic: Raynaud's disease (1960–1985).

Topic: Raynaud's, limited to publications before 1986
PubMed search: Raynaud AND human AND 1960[DP]:1985[DP]
Number of documents retrieved: 2,733
Number of MeSH term instances in the document set: 52,271
Number of unique MeSH terms in the document set: 2,972
Profile: (top 5 terms for a few semantic types are shown below)
Semantic Type: Body Space or Junction: {finger joint (1.0), wrist joint (0.81), elbow joint (0.55), esophagogastric junction (0.33)}
Semantic Type: Cell: {neutrophils (1.0), blood platelets (0.78), erythrocytes (0.71), eosinophils (0.53), lymphocytes (0.5)}
Semantic Type: Cell Function: {platelet aggregation (1.0), platelet adhesiveness (0.56), neural conduction (0.5), erythrocyte aggregation (0.44)}
Semantic Type: Organ or Tissue Function: {regional blood flow (1.0), microcirculation (0.41), vasoconstriction (0.41), blood flow velocity (0.41), hemodynamics (0.31)}
Semantic Type: Disease or Syndrome: {Raynaud's disease (1.0), scleroderma, systemic (0.23), vascular diseases (0.09), occupational diseases (0.077), cold (0.074)}
Semantic Type: Eicosanoid: {epoprostenol (1.0), prostaglandinse (0.65), prostaglandins (0.52), alprostadil (0.51), prostaglandinse, synthetic (0.15)}
Semantic Type: Organism Function: {aged (1.0), blood pressure (0.29), exertion (0.1), body temperature regulation (0.09), pregnancy (0.07), menstruation (0.04)}
Semantic Type: Physiologic Function: {blood viscosity (1.0), blood circulation (0.63), pulse (0.38), vascular resistance (0.33), collateral circulation (0.13)}
Number of unique MeSH terms in profile: 2,972
Total number of MeSH term entries in profile: 4,419 (a term can be in multiple semantic types)
Top 5 Semantic types ranked by number of terms: Disease or Syndrome (686), Pharmacologic Substance (359), Organic Chemical (291), Laboratory Procedure (224), Body Part, Organ, or Organ Component (198)
Number of semantic types with at least 1 term in profile: 114 (out of 134 possible)

Note. Only the top 5 MeSH terms and weights are shown within select semantic types.

may be limited to functional semantic types such as *Cell Function* and *Pathologic Function*. In a recent article (Srinivasan & Wedemeyer, 2003), we used profiles of diseases limited to *Geographic Area* to explore the global distribution of research on various diseases. We then compared such distributions with disease prevalence (distribution) data. In this research, our goal is to employ MeSH-based profiles for hypothesis discovery.

Example Profile

We use "Raynaud's disease" to illustrate topic profiles. Table 1 presents details of the search performed, counts pertaining to the set of retrieved documents, and the profile. Five top-ranked MeSH terms and their weights (equation 3) are shown for a sample set of semantic types. Semantic types with the most terms are also identified. For example, *Disease or Syndrome* with 686 terms is the top-ranked type. With respect to the distribution of weights, a threshold of 0.5 yields a profile of 316 terms, which is only 7% of the original 4,419 terms. A threshold of 0.3 gives 413 (9.3%) while a threshold of 0.7 gives 162 terms, which is less than 1% of the terms.

Discovering Hypotheses

Consider now a user who is interested in a particular disease. Perhaps she wants to identify genes that may be associated with this disease or dietary factors that influence the disease in some way. The kinds of connections of interest here are those that are both indirect and novel. The discovery process initiated by Swanson explores such connections and may be described using Figure 3. If we focus first on Figure 3 (left) our user's disease of interest is represented by topic A. Links through intermediate terms (B1, B2, . . .) lead to the topics represented by terms C1, C2, etc. By implication, there may be an interesting indirect association between A and C1, A and C2, etc., via the linking B terms. An association is novel (say between A and Cx) if the two have been studied independently, i.e., their literatures do not overlap. The goal in the discovery process is to automatically identify such C level concepts given the user's starting A topic. Note that A can represent any topic such as a disease or a gene or an enzyme or a pharmacological substance. Observe that the B terms have an important role because these represent conceptual bridges between A and C.

Instead of an A to B to C discovery approach, a user may start with a pair of topics A and C. While it may be that some connections between them are already known, the aim is to find new ones. Alternatively, it may be that no connections are known and the aim is to determine if a meaningful connection is possible. This time referring to Figure 3 (right), the bi-directional process starting from both A and C looks for novel and meaningful interlinking B terms. This is, in fact, that discovery pathway used by Swanson (1986) for his first discovery. Specifically, he was able to identify from MEDLINE mechanisms supporting his intuition that Raynaud's disease (A topic) may be treated with fish oils (C topic). He found that Raynaud's is aggravated by high blood viscosity, platelet aggregability, and vasoconstriction and these are reduced by fish oils. Weeber et al. (2001) in their replication of some of the work by Swanson and Smalheiser, labeled the one directional procedure as an "open"

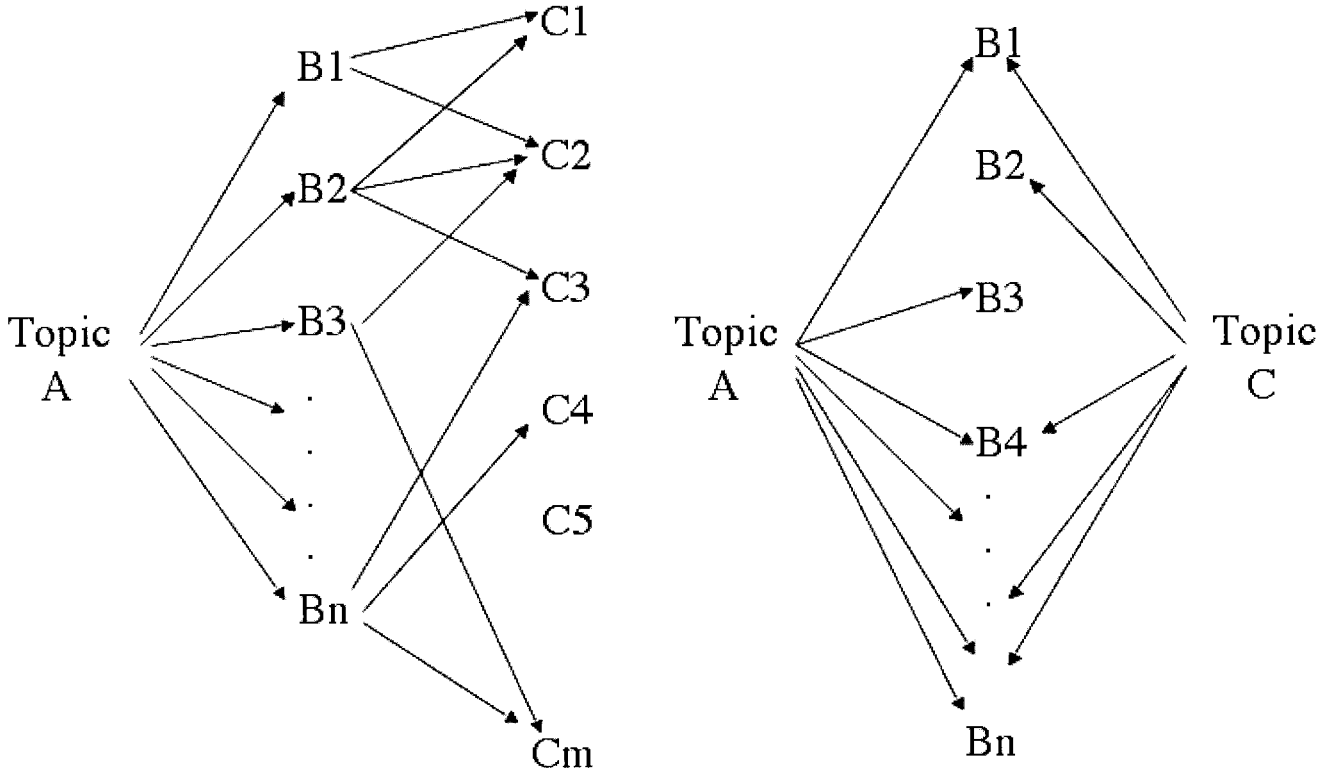

FIG. 3. Indirect concept links. (Figure adapted from Weeber et al., 2001.)

discovery procedure and the bi-directional one as a "closed" discovery process. We adopt the same labels henceforth.

The above description of the open and closed discovery approaches is general enough that it may be applied to any subject domain and not just those represented by the MEDLINE database. The key is to instantiate the nodes, arcs, and pathways in appropriate ways. In parallel research, we are exploring their application to the accounting domain.

Several key aspects need to be considered when implementing a system to support the open and closed discovery processes. It is in the decisions made regarding these aspects that the various studies may be distinguished. First, how are the A topics for open (A and C for closed) to be represented? In our work, these are essentially MEDLINE searches that one would normally submit to the PubMed search system. Next, what are the intermediate B terms and how are these identified? Similarly, how are the C terms identified in the open process? For example, in our open process B terms are MeSH terms identified from documents retrieved from the A search. These are then filtered and select B terms are retained. C terms, also MeSH terms, are then identified from documents retrieved from these B term searches. Instead of MeSH, others have explored extracting phrases from the free-text portions of MEDLINE documents. Another decision that must be made is regarding constraints that may be usefully applied to the processes. For example, in the open process an explosive number of B terms and then C terms can occur. We use term weights and also the UMLS semantic types as constraints. Top-weighted terms within specific semantic types are selected. Others have generally followed a manual term selection phase sometimes in conjunction with similar (but not identical) term weight–based constraints.

Another important aspect of the discovery process is regarding the nature of the relationship represented by the arrows in the figures. This is where the domain expert has the greatest influence over the process. Using background knowledge, the expert is expected to filter out and select the most promising paths for exploration. In this way, the expert user is intimately involved with the discovery processes. We support the user in this filtering step by providing mechanisms to look at the literature underlying each arc. For example, underlying the arrow between A and B2 in Figure 3 (left) is the set of documents retrieved by an intersection of the A topic search and the B2 MeSH term. Others provide similar support by displaying appropriate sentences from the records. In summary, when we examine key details, we see that text mining methods vary significantly although designed with the common goal of enabling open and closed discoveries.

Swanson and Smalheiser made most of their predictions using the closed process. For example, Smalheiser & Swanson (1996a) predicted connections between indomethacin and Alzheimer's disease and Swanson (1988) predicted 11 different pathways between migraine and magnesium. The challenge in their discovery process is that it requires considerable manual processing. They start with independent literature searches on the initiating A and C topics. Titles of

retrieved documents (or terms extracted automatically from them) are compared manually for interesting potential connections between A and C. In the migraine (A) – magnesium (C) problem, Swanson observed, for example, that magnesium deficits can lead to high levels of serotonin release and substance P activity. He also observed from the A literature that these same phenomena tend to aggravate vascular effects of migraine. Thus, by reading the titles in these two literatures and focussing on common terms such as serotonin and substance P, Swanson was able to suggest that magesium may have a role in migraine.

More recently, Swanson and Smalheiser created ARROWSMITH (Swanson & Smalheiser, 1997), a system available for public use on the WWW[3] that is designed to assist with the discovery process. Functions provided include those to automatically extract words and phrases that are in common between the two A and C document sets. Although the tool is extremely helpful, the process still involves significant human intervention in sifting through the list of terms and selecting appropriate terms.

To the best of our knowledge, Gordon and Lindsay were the first researchers to try to replicate the Swanson discoveries. They explored both the Raynaud's–fish oils problem (Gordon & Lindsay, 1996) and the migraine-magnesium problem (Lindsay & Gordon, 1999). Although their strategy parallels the one used by Swanson and Smalheiser, there are several distinctive features. They extract terms (bigrams and trigrams) from the free-text portions of records and assess their potential value using 4 different weighting schemes. These are: term frequency, record frequency, term frequency* inverse document frequency where the latter is computed against all of MEDLINE, and, finally, record frequency in the domain normalized by record frequency in all of MEDLINE. In their early work, they combine the evidence obtained from the four weighting schemes to get a final ranking of the terms while later they concentrate more on using relative record frequency. Term selections from ranked lists are done manually as are the design of search strategies.

Weeber et al. (2000, 2001) in their knowledge discovery research use MetaMap (Aronson & Rindflesch, 1997), an NLP system, to translate the MEDLINE free-text (titles and abstracts) to terms from the UMLS (Unified Medical Language System) vocabulary. There are several advantages to working with UMLS terms instead of ngrams extracted from free-text. For example, given the domain of the UMLS, terms are more likely to be biomedically relevant. Their procedures also emphasize co-occurrence at the sentence level and semantic filters. For example, referring to the left part of Figure 3, given a particular A topic, B terms selected are UMLS terms of particular semantic type(s) that co-occur with terms representing topic A at the sentence level. Although the semantic filters drastically limit the size of the extracted term set, their discovery process still involves significant manual analysis to group terms representing pathways and select terms for search strategies. It may be observed that enforcing sentence-based co-occurrence potentially constrains complexity of the input topic in the discovery system.

As mentioned briefly before, we view the discovery process as having two dimensions. The first is the identification of key terms (find B then C in the open discovery process, for example). The second is determining the nature of the interlinking relationship (between A and B or between B and C). We call these dimensions and not stages since a given discovery procedure may intertwine them in complex ways. The second dimension at least benefits from, if not depends upon, input from a domain expert. This process may, of course, be assisted by automated mechanisms as, for example, the convenient display of contextual sentences (Weeber et al., 2001). Our focus in this research is on dimension 1 and we wish to explore the value of MeSH metadata for term selection. Specifically, our goal is to provide a usefully ranked list of MeSH terms to the end user. In the open discovery process, we wish to provide ranked C terms and in the closed process ranked B terms.

Set in this research context, we now present our open and closed algorithms. Our approach is similar to the research of Lindsay and Gordon in that we use term weights that go beyond simple frequency counts. It is similar to the efforts of Weeber et al., in that we use UMLS-based semantic filters. It is distinctive in that we use MeSH-based profiles. Moreover, the user has only to specify the kinds of profiles to build for the problem and set a parameter. Other characteristics of our methods are identified later.

Open Discovery Algorithm

Figure 4 outlines the various steps our open discovery algorithm. First, a MeSH profile is built for the initiating A topic. MeSH terms in the profile have TF*IDF weights that are normalized within each semantic type (equation 3). *N* B MeSH terms are automatically selected from the user-specified ST-B vector components and their profiles are in turn built in step 3. These are then merged in step 4 to form a final combination profile. The combined weight of a term is the sum of its weights in the individual B profiles. In the last step, the C MeSH terms are limited to those representing novel connections. A C MeSH term's score is regarded as the system-derived estimate of the potential value in its association with the A topic. This score depends both on the number of paths connecting back to A as well as the strengths of these paths. The higher the score, the stronger the recommendation made by the system. Thus, we rank C MeSH terms within each semantic type by its combined weight.

The input A topic may represent any topic of interest to the user. The search that is conducted to build the A profile may be any appropriate PubMed search and need not be limited to the metadata field or indeed consist of any meta-

[3] There are two implementations available on the Web, which are at http://kiwi.uchicago.edu and http://arrowsmith.psych.uic.edu.

Input from user: (1) an A topic of interest, (2) a set of UMLS semantic types of interest (ST-B) for selecting B terms and a set (ST-C) for selecting C terms.
Parameter: N

- Step 1: Conduct an appropriate PubMed search for topic A, and build its MeSH profile limited to the semantic types in ST-B (equation 3). Call this profile AP.
- Step 2: For each semantic type in ST-B, select the N top ranking MeSH terms from AP. These are designated the B terms (B1, B2, B3, etc.).
- Step 3: Conduct an independent PubMed search for each B term and build its profile limited to the semantic types ST-C (equation 3). Call these profiles BP1, BP2, BP3, etc.
- Step 4: Compute a final combined profile where the combined weight of a MeSH term is the sum of its weights in BP1, BP2, BP3, etc. Call this initial profile CP.
- Step 5: For each term t in CP if a MEDLINE search on topic A AND t returns non zero results, eliminate t from CP.

Output: For each semantic type in ST-C, output the MeSH terms in CP ranked by combined weight. These are the C terms organized by semantic type and ranked by estimated potential.

FIG. 4. Open discovery algorithm: Outline of steps.

data search terms. Thus, the A topic may be as complex as required by the user. Our current strategy for searching is to take the user's input directly as search terms and add the constraint of limiting the retrieved records to human studies.

If the user is uncertain about how to specify ST-B and ST-C, these may be left unspecified. The default strategy is to use all available semantic types. However, this default strategy is likely to produce more ambiguous results than the situation where the user actually specifies types of interest. Also, the parameter N controls the width of the expansion from B to C. Smaller values are likely to yield more focussed output. We present experiments exploring variations in N as well as in the semantic types.

Closed Discovery Algorithm

Figure 5 outlines the key steps in this algorithm. Profiles are built for topics A and C in the usual way but limited to

Input from user: (1) Two topics of interest designated, A and C and (2) semantic types of interest (ST-B) for selecting intermediate B terms.
Parameter: P.

- Step 1: Conduct independent PubMed searches for A and C. Build the A and C MeSH profiles (equation 3). Call these profiles AP and CP respectively.
- Step 2: Limit AP and CP to component vectors for semantic types listed in ST-B. Also for each ST-B semantic type retain only the highest weighted P MeSH terms within AP and CP.
- Step 3: Compute a B profile (BP) composed of terms in common between AP and CP. The weight of a MeSH term in BP is the sum of its weights in AP and CP.
- Step 4: Eliminate terms from BP that do not represent novel associations. That is, if a search for A AND t AND C returns non zero results, then eliminate t from BP. The remaining MeSH terms are the potential B concepts of interest. (This constraint may be relaxed if the user is also interested in connections that are not necessarily novel).

Output: For each semantic type in ST-B, display a ranked list of MeSH terms to the user. Each term represents a potential conceptual connection between A and C.

FIG. 5. Closed discovery algorithm: Outline of steps.

the ST-B semantic types. By specifying ST-B, the user is in essence suggesting that terms belonging to these semantic types may potentially link A and C. Again, the A and C topics may be as complex as the user requires. *P* top ranking terms are retained for each semantic type. A profile of MeSH terms in common between AP and CP is then built that represents potentially novel connections. If the user is unable to specify semantic types for selecting B terms, then the default strategy will be to display rankings within all available semantic types. Also, the constraint in step 4 may be relaxed to offer the user the opportunity to explore connections between A and C that may not be totally novel in the literature.

Summary

Our open and closed discovery processes based on MeSH profiles are guided by the semantic types identified by the user. For certain kinds of problems, particular semantic types may be most suitable. For instance, when focussing on diseases, the functional semantic types appear most relevant. Weeber et al. (2000, 2001) also use semantic types to limit the terms considered. However, we also display the output terms grouped and ranked within semantic type while their ranking is across all filtered terms independent of semantic types. Grouping by semantic type is, in fact, built into our procedures. Lindsay and Gordon differentiate between single words and higher order ngrams, but make no attempt to partition terms semantically. We believe that semantic groups will be more useful to the end user since entire groups may be eliminated (or focussed upon) with greater ease.

In general, our open and closed discovery algorithms are designed as the foundation of our text mining system. Our larger goal is to offer users, i.e., domain experts, a suite of text mining tools that supports the exploration of MEDLINE for the purpose of hypothesis generation. Our aim is to automate the process as far as possible. However, user input on several decisions will be key to successful knowledge discovery.

Discovery Problems

Our goal is to assess the effectiveness of our open and closed discovery algorithms. The question we ask is: can our MeSH-based text mining methods identify interesting hypotheses? In response, we test the ability of our methods to replicate the various discoveries made by Swanson and Smalheiser. Weeber et al. (2000, 2001) and Lindsay and Gordon (1996, 1999) use the same empirical strategy to test their discovery methods. As discussed previously, we focus on the first dimension of the discovery process, i.e., on identifying potentially interesting terms related to the input topic(s). Designing strategies to assist in the second dimension, i.e., when identifying the specific nature of the relationship, is left for the future. Thus, in our simulation experiments our aim is to see if we can automatically identify the key terms that are at the core of each of the Swanson and Smalheiser discoveries. For each discovery, we conduct open or closed discovery runs as relevant to the problem. As seen in Figures 4 and 5, our algorithms generate ranked lists of terms within each semantic type. We measure performance by the ranks of the key MeSH terms (C or B terms) within the appropriate semantic type for each problem. For each problem that we replicate, we first summarize the original discovery, then present our results and then, where available, we also summarize experiments by others.

Fish Oils and Raynaud's Disease

The Raynaud's discovery problem (Swanson, 1986) was introduced earlier Swanson observed that Raynaud's is exacerbated by platelet aggregability, vasoconstriction, and blood viscosity. He also observed from the literature that fish oils reduce these phenomena. Putting these two observations together, he postulated that fish oils may be beneficial for persons with Raynaud's. This was later confirmed (DiGiacomo et al., 1989). Although the original discovery was made using the closed approach, we first apply the open procedure (Gordon & Lindsay, 1996; Weeber et al., 2001) and then the closed one (Weeber et al., 2001).

Open discovery. Raynaud's disease as the A topic initiates our open discovery algorithm (Fig. 4). The Raynaud's search is limited to human studies in the time period 1960 to 1985 (since the discovery was made in 1986). The Raynaud's profile is limited to a view defined by eight functional semantic types (ST-B): *Cell Function, Finding, Molecular Function, Organism Function, Organ or Tissue Function, Pathologic Function, Phenomenon or Process,* and *Physiologic Function*. *N* is set to 1 in order to keep the discovery process focused. The selected (top ranking) B terms in the order of the semantic types listed above are: *platelet aggregation; scleroderma, systemic; antibody specificity; aged; regional blood flow; thrombosis; recurrence,* and *blood viscosity*. Eight B profiles are built for these terms from the same publication time period again constrained to human studies. These are then merged to obtain the combined C profile. Terms within each semantic type are ranked by combined weight. Were a real user present, these ranked lists would be shown to this user who may then select particular semantic types to browse. Since Swanson was interested in dietary factors, the appropriate semantic types for the C terms would be *Element, Ion, or Isotope*; *Vitamin*; and *Lipid*. Out of these three, the results are most interesting for the *Lipid* type. Table 2 (R1 column) lists the ranks of the top 20 terms that are novel in that they do not co-occur with Raynaud's within the pre-1986 MEDLINE database (step 5 of algorithm).

5,8,11,14,17-eicosapentaenoic acid is an important polyunsaturated fatty acid found in fish oils. It is also the active ingredient in it. Thus, we see that the concept of fish oils is

TABLE 2. Raynaud's (open discovery).

MeSH term	R1	R2	R3	MeSH term	R1	R2	R3
Lipoproteins, LDL	1	1	1	Glycolipids	11	8	11
Oils	2	3	3	Chylomicrons	12	13	17
Lipoproteins, HDL	3	2	2	Linolenic acids	13	20	—
Lipoproteins, VLDL	4	5	5	Glycerides	14	13	—
Platelet activating factor	5	6	13	Lipid peroxides	15	10	8
Phosphatidylcholines	6	4	4	Butyrates	16	14	16
5,8,11,14,17-eicosapentaenoic	7	17	27	Sodium tetradecyl sulfate	17	25	18
Lysophosphatidylcholines	8	7	18	Cardiolipins	18	20	22
Gangliosides	9	12	9	Fish oils	19	28	46
Iodized oil	10	15	14	Liposomes	20	11	7

Note. R1: N = 1 and ST-B = 8 functional semantic types, R2: N = 2 and ST-B = 8 functional semantic types, R3: N = 1 and ST-B = all 134 semantic types. Cell values are ranks. Terms listed are top 20 C MeSH Terms in semantic type *Lipid* for R1.

indicated within rank 7. The MeSH term *fish oils* itself is ranked 19th. The other entries are also meaningful. For example, lipoproteins have been studied in association with fish oils as seen, for instance, in a 1985 document that indicates a reduction of plasma lipids and lipoproteins by marine fish oils.[4] To complete the logical connection, oxidation of low-density lipoproteins has been recently studied in conjunction with Raynaud's.[5] Also, a 1999 study explores the effect of Probucol, an antioxidant that influences low-density lipoprotein oxidation time lag, on Raynaud's.[6]

We chose 8 functional semantic types to select the intermediate B terms. The assumption made is that the user has an interest in factors that influence the functional aspects of the disease. These categories were also used in replications of the same open discovery by Weeber et al. (2000, 2001). We observe that after setting $N=1$ and ST-B to the 8 functional semantic types, our procedure automatically identifies the key term at a high rank. This is despite the fact that some of the intermediate B terms such as *recurrence* and *aged* are very general. The remaining columns of Table 2 show rankings when we vary the parameters. R2 shows the ranks of terms when $N=2$. (Note that ranks are shown only for those top 20 terms from R1.) For most of the terms, the ranks are lowered. The key terms *5,8,11,14,17-eicosapentaenoic acid* and *Fish Oils* are now at ranks 17 and 28, respectively. The R3 column represents ranks when using $N = 1$ but all 134 semantic types for ST-B. Interestingly, although the ranks for *5,8,11,14,17-eicosapentaenoic acid* and *Fish Oils* drop to 27 and 46, respectively, the ranks of the other terms do not change dramatically. Guided by the ranks of the key terms, we conclude that it is desirable to set N to 1 and define ST-B as appropriate for the problem. Using the same criteria, it appears that the R3 run is less effective than R2.

Weeber et al. (2001) in their UMLS concept-based approach identified from the Raynaud's literature a set of 145 B concepts that belong to the same eight functional semantic types. They manually analyzed these and identified three groups of concepts representing three pathways. Each group was then explored independently. (In contrast, we select the highest ranked B MeSH terms and conduct searches automatically.) For the two groups representing platelet aggregation and blood viscosity, their best ranks obtained for concepts relevant to fish oils were 50 and 20, respectively. The third pathway for vascular reactivity did not produce any reasonable ranks for fish oil concepts.

Gordon and Lindsay (1996) extracted single words and bigrams from the free-text portions of the Raynaud's literature. After manual analysis, they recognized that blood was an important aspect. They then hand-selected a group of relevant terms such as capillary abnormalities and digital artery. Manual examination of the literature pertaining to the intersection between these blood terms and Raynaud's led them to blood viscosity. Documents on blood viscosity yielded several lists of terms from which they hand-picked 115 interesting terms. These were hand-picked from the top 250 terms obtained from each of their 4 term weighting schemes applied independently to single-and multi-word phrases. After eliminating terms from the set of 115 that already occurred with Raynaud's, they were left with 34 terms that contained both fish oils and eicosapentaenoic acid. However, their ranks are not provided.

Closed discovery. Here we assume that the mechanisms of the interaction are unknown and this is our discovery goal. We know that Swanson identified platelet aggregation, blood viscosity, and vasoconstriction as representing the key connections. So the question here is can we rediscover these as the connecting B MeSH terms? We use the procedure in Figure 5 with Raynaud's as topic A, and the key fish oil concept: 5,8,11,14,17-Eicosapentaenoic Acid as topic C.

Profiles for A and C are created from the pre-1986 human studies and limited to the functional semantic types (ST-B). Parameter P is set to 10 or 20 or 30 or 40 when generating the combined profile. Table 3 showing the ranks of key B MeSH terms indicates that all three pathways are

[4] PMID: 3903563, a 1985 publication.

[5] PMID: 7639801, a 1995 publication.

[6] PMID: 10378706, a 1999 publication.

TABLE 3. Terms representing pathways between Raynaud's and fish oils.

MeSH term	Semantic type	Term rank			
		P = 10	P = 20	P = 30	P = 40
Platelet aggregation	Cell function	1	1	1	1
Platelet adhesiveness	Cell function	2	2	2	2
Blood viscosity	Physiologic function	1	1	1	1
Vasoconstriction	Finding	1	1	1	1
Thrombosis	Finding	2	2	2	2

represented within the top 3 ranks. We also observe that changing the value for *P* has no effect on the ranks (We see later that this is not always the case.)

We also repeated the $P = 20$ run, but with different searches for the C topic. Searching on (5,8,11,14,17-Eicosapentaenoic Acid OR fish oils) yields slight changes: *vasoconstriction* is ranked 4 and *thromobosis* is ranked 3. The same results are obtained when C is represented by the search (fish oils). This is not surprising since these variant searches yield almost the same document set with the former retrieving 7 additional documents for a total of 578 documents. In contrast, the original run with 5,8,11,14,17-Eicosapentaenoic Acid as the search retrieves only 207 documents. Despite the differences in retrieved sets, the conclusions are almost identical.

Exploring the same problem under the closed discovery approach, Weeber et al. (2001) state that they found the appropriate connecting concepts. Unfortunately, the ranks of these intermediate concepts are not stated. Hence, we cannot compare our results.

Migraine and Magnesium

Swanson (1988) studied the problem of finding connections between migraine and magnesium. In this case, he studied both the titles and MeSH terms of MEDLINE records. Using his method, he was able to propose 11 neglected paths that potentially connect migraine with magnesium. Gallai et al. (1992), for example, were later able to corroborate these connections. Swanson observed, for example, that magnesium deficits can lead to high levels of serotonin release and substance P activity. These, in turn, tend to aggravate vascular effects of migraine.

We now consider the migraine-magnesium problem in both open and closed discovery modes. One observation to make here is that the open discovery is a bit forced since by 1988 there were already a few documents in which migraine and magnesium co-occur. Thus, the two literatures are slightly connected. Hence, when applying our open discovery algorithm, we will have to modify step 5 in which we remove concepts that are already co-occurring with the A concept. Instead, we apply a threshold for the co-occurence frequencies. Similar modifications were made by Lindsay and Gordon (1999) and Weeber et al. (2001) in their replications of this discovery problem.

Open discovery. We start our open discovery process with the pre-1988 MEDLINE subset and migraine as the A concept. The same group of eight functional semantic types from the Raynaud's problem are used for selecting intermediate B concepts and we set $N = 1$. The top-ranking B concepts chosen were: *platelet aggregation* in *Cell Function; contraceptives* in *Finding; drug interactions* in *Molecular Function; aged* in *Organism Function; cerebrovascular circulation* in *Organ or Tissue Function; cerebrovascular disorders* in *Pathologic Function, recurrence* in *Phenomenon or Process*; and *pulse* in *Physiologic Function.*

These B concepts finally lead to a C profile that was analysed for magnesium. Since there are already six documents in which migraine and magnesium co-occur within the 1988 MEDLINE records, we keep 10 as our co-occurrence limit. We then find that *magnesium* is in rank 5 under the metadata category *Element, Ion or Isotope.* We have *technetium, iodine radioisotopes, tritium, cobalt radioisotopes* in ranks 1 through 4 before *magnesium*. Looking at co-occurrences in MEDLINE records until 2003, we find there are 45, 23, 8, 0, and 48 documents in which migraine co-occurs with these concepts, respectively. *Technetium* and *iodine radioisotopes*, for example, are involved in a tomography procedure used in studying migraine.

Lindsay and Gordon (1999) started with searches representing the intermediate literatures and explored these pathways to determine if magnesium could be found. More specifically, using just the relative frequency weighting strategy, they formed the union of the top concepts from different intermediate pathways. Deleting those that were also found using all four statistics, they were left with 34 concepts that also included magnesium. Unfortunately, the deletion process is not clearly justified. If we ignore this deletion step, then magnesium is found within a ranking of 80 concepts. Their procedure is also not strictly an open process since they start with the intermediate pathways.

Weeber et al. (2001) analysed MEDLINE for migraine and extracted over 3,000 concepts that co-occurred with migraine in sentences. The functional semantic types filter reduced these to 504 concepts. Four pathways were recognized and followed with further analysis. Concepts were hand selected for these pathways. Analysing the literatures of these pathways, they find that magnesium appears in the

TABLE 4. Terms representing pathways between migraine and magnesium.

MeSH term	Semantic type	Term rank			
		P = 10	P = 20	P = 30	P = 40
Vascular mechanisms					
Membrane potentials	Cell function	1	1	1	1
Biological transport	Cell function	2	2	2	2
Cell membrane permeability	Cell function	3	3	3	3
Action potentials	Cell function	4	4	4	4
Vascular resistance	Laboratory or test result	3	5	5	5
Spreading depression					
Spreading cortical depression	Laboratory or test result	—	6	6	6
Depression	Mental or Behavioral dysfunction	—	4	4	5
Prostaglandins					
Prostaglandins	Eicosanoid	1	1	1	1
Prostaglandins E	Eicosanoid	2	2	2	2
Arachidonic acids	Eicosanoid	—	3	3	3
Prostaglandins F	Eicosanoid	3	4	4	4
Alprostadil	Eicosanoid	4	5	5	5
Epoprostenol	Eicosanoid	5	6	6	6
Serotonin					
Serotonin	Neuroreactive Substance or biogenic amine	1	1	1	1
Platelet activity					
Platelet aggregation	Cell function	—	6	6	6
Platelet aggregation	Laboratory or test result	2	2	3	3
Platelet adhesiveness	Cell function	—	11	11	11
Calcium channel blockers					
Calcium channel blockers	Injury or poisoning	—	3	4	4
Magnesium sulfate	Inorganic chemical	—	3	3	3
Type A personality					
Anxiety	Mental process	1	1	1	1
Stress	Pathologic function	2	2	4	4
Personality	Mental process	—	2	4	4
Aggression	Social behavior	—	1	2	2
Inflammation					
Edema	Pathologic function	—	2	6	9
Brain edema	Pathologic function	—	—	—	10
Inflammation	Pathologic function	—	11	10	15
Hypoxia					
Anoxia	Pathologic function	—	—	5	6
Epilepsy		—	—	—	—
Substance P					
Substance P	Neuroreactive Substance or biogenic amine	—	18	19	20

Note. A dash implies that the term was not found.

top 15% of the ranked lists of C concepts for 3 of the pathways. The actual ranks appear to be between 10 and 42.

The previous research indicates that finding magnesium when starting from migraine is a challenging problem. Despite this, we see that our open discovery process is able to successfully rank magnesium high (rank = 5). Moreover, it does so automatically after N and semantic types for ST-B are specified. Repeating the same experiment with $N = 2$ and keeping everything else the same yields a rank of 12 for magnesium. Interestingly, keeping $N = 1$ but using all semantic types for ST-B yields a rank of 5 for magnesium. Thus, the rank of this key term is preserved even when we use all semantic types for selecting B terms. This is encouraging since at least in this particular case performance does not depend upon specifying intermediate semantic types for the open discovery process.

Closed discovery. For the closed procedure, we work with migraine as the A topic and magnesium as the C topic and limit the process to pre-1988 MEDLINE. We leave ST-B unspecified and instead look for the ranks of key terms within appropriate semantic types. (We adopt the same strategy in the remaining closed discovery problems.) Table 4 identifies the 11 connections (such as vascular mechanisms) between migraine and magnesium that Swanson proposed. If we consider the P = 20 column, we see that eight of the 11 connections are visible using our method within the top 6 ranks, with vascular mechanisms and prostaglandins being very well represented. In fact, the majority of the MeSH terms under *Eicosanoid* are prostaglandins some of which are shown in Table 4. Amongst the remaining eicosanoids, we have, for example, *arachidonic acids,* which is an unsaturated, essential fatty acid and a precursor

TABLE 5. Terms representing pathways between Alzheimers and indomethacin.

MeSH term	Semantic type	Term rank			
		P = 10	P = 20	P = 30	P = 40
Signal transduction	Molecular function	2	4	5	5
Lipid peroxidation	Molecular function	4	6	7	7
Membrane fluidity	Molecular function	—	10	11	11
Oxidative phosphorylation	Molecular function	—	14	15	15
Receptors, muscarinic	Receptor	3	5	5	5
T-lymphocytes	Cell	—	5	6	6
Lymphocytes	Cell	4	7	8	8
Acetylcholine	Neuroreactive sub. or biogenic amine	2	2	2	3
Thyrotropin	Neuroreactive sub. or biogenic amine	—	9	12	13

Note. A dash implies that the term was not found.

in the biosynthesis of prostaglandins. Additional observations may be of interest. A 1997 review[7] states that abnormalities in catecholamines and endogenous opioids are part of the biomechanisms involved in migraine. Interestingly, in this pre-1988 analysis we have several cetecholamines ranked high under *Neuroreactive Substance or Biogenic Amine* (after *serotonin*). *Catecholamines* itself is at rank 6. Particular catecholamines such as *norepinephrine, epinephrine,* and *dopamine* are at rank 2, 4, and 8, respectively. There are also key connections between entries under different semantic types. For example, *epoprostenol* (rank 6) in *Eicosanoid* is a prostaglandin, a powerful vasodilator, and it inhibits *platelet aggregation*.[8] Unfortunately, one pathway (epilepsy) was not identified by our algorithm. Also, a second pathway substance P was ranked rather low.

Analysing the effect of parameter *P,* we see that although the ranks obtained are slightly better using P = 10, several key MeSH terms are not found. In fact, only 5 of the pathways are represented. We note that by increasing the parameter values beyond 20, we do not get significant changes. Although the ranks change slightly for a few terms, on the whole the ranks show remarkable stability.

In their closed discovery process, Weeber et al. (2001) find 253 functional concepts that co-occur with both migraine and magnesium. Examining this list, they find 6 of the 11 pathways discovered between migraine and magnesium. Again, we are unable to compare our results since the ranks of these pathway concepts are not provided.

Indomethacin and Alzheimer's Disease

Smalheiser and Swanson (1996a) explored possible mechanisms by which indomethacin, an anti-inflammatory agent, might be expected to affect patients with Alzheimer's. This discovery was made using the MEDLINE literature limited to June 1995. They made several observations. For example, they state "Indomethacin decreases plasma membrane fluidity in various cell types, whereas membrane fluidity is elevated in some patients with AD." Membrane fluidity is the connecting concept in this statement. Similarly they observed connections pertaining to killer T-cell activity, M2-muscarinic receptors, lipid peroxidation, and thyrotropin-releasing hormone.

Table 5 presents the results obtained from our closed discovery process with Alzheimer's disease and indomethacin as the A and C topics and their profiles derived from pre-June 1995 and "human" studies in MEDLINE. Parameter *P* is varied as before. Again we see that P = 20 provides the best results with all key terms but *oxidative phosphorylation* ranked within the top 10 positions. Note that "T-cell" translates to the MeSH term *t-lymphocytes.* The ranks are somewhat better than reported if we decide to eliminate very general terms from the top ranks. For example, we have *drug interactions; stimulation, chemical;* and *enzyme activation* as the top 3 concepts within *Molecular Function* (not shown).

Interestingly, Smalheiser and Swanson also discovered a possible adverse effect related to indomethacin inhibition of acetylcholine in several systems such as smooth muscles. We note that *acetylcholine* is ranked second under *Neuroreactive Substance or Biogenic Amine.* This example emphasizes the nature of text-based knowledge discovery. Although the key MeSH terms may be ranked high, it is left to the user to figure out that indomethacin *decreases* membrane fluidity or that it *inhibits* lipid peroxidation. It is also left to the user to differentiate between positive connections and those that represent adverse effects (as in the case of acetylcholine). We hope that by focusing on the top-ranked terms identified by our system, the user will have to peruse far fewer documents in order to obtain the details regarding the nature of the mechanism linking A and C.

One further observation may be made related to *nitric oxide,* which is ranked first under *Inorganic Chemical.* Several pre-1996 reports identified nitric oxide as important for understanding Alzheimer's. Nitric oxide synthase activity is elevated in brain microvessels in Alzheimer's. Elevated vascular production of NO may contribute to the

[7] PMID:9100398, 1997.

[8] PMID:6988063, 1980.

TABLE 6. Terms representing pathways between somatomedin C and arginine.

MeSH term	Semantic type	Term rank			
		P = 10	P = 20	P = 30	P = 40
Lymphocytes	Cell	—	6	7	7
T-lymphocytes	Cell	—	10	13	13
Cell division	Cell function	1	1	1	1
Cell differentiation	Cell function		3	3	3
Cell survival	Cell function	4	4	4	4
Lymphocyte transformation	Cell function	5	5	5	5
Somatotropin	Hormone	1	1	1	1
Somatotropin-releasing hormone	Hormone	—	—	4	5
Somatostatin	Hormone	3	3	5	6
Wounds and injuries	Injury or poisoning	—	4	4	4
Body weight	Organism attribute	4	4	4	4

Note. A dash implies that the term was not found.

susceptibility of neurons to injury and cell death in Alzheimer's.[9] On the other side, we also find that indomethacin prevents induction of nitric oxide synthase.[10] Interestingly, since 1995 the connection via nitric oxide has been studied further. In 2000, for example, evidence was obtained showing that indomethacin reduces interferon-gamma-induced NO production. Accompanied by an inhibition of inducible nitric oxide synthase mRNA expression.[11] Also, a 2001 study further examines the neuroprotective characteristics of indomethacin.[12]

Somatomedin C and Arginine

Swanson (1990) explored the relationship between somatomedin C (which we refer to as SmC), a growth-regulating peptide believed to be mainly active in adults, and arginine, an essential amino acid. His analysis of the two isolated literatures revealed several connections between them. We mention a few here and refer the reader to Swanson's report for the details.

Growth hormone (GH) and SmC influence each other while arginine stimulates the secretion of GH. SmC levels and GH secretion decline with age. Arginine is effective in treating emaciation in older individuals and the general decline of lean body mass. Its level is also reduced in patients with protein-calorie malnutrition. SmC promotes wound healing after burns, NK-cell activity, and immune functions. Arginine does the same.

Our closed discovery process on the pre-1990 literature on human studies brings up quite a few top-ranked terms related to cell growth, which is key to the aging process. As Table 6 shows, these include for example *cell division* (rank 1) under *Cell Function.* Related to the notion of wounds we have *wounds and injuries* (rank 3) under *Injury or Poisoning.* Although we don't have entries for NK-cell (natural killer cell) activity, we do have entries for its parent type *lymphocytes* (rank 6) and related terms *t-lymphocytes* (rank 10) under *Cell* as well as *lymphocyte transformation* (rank 5) under *Cell Function.* Under *Hormone*, we have at rank 1 *somatotropin* (Growth Hormone) and *somatostatin* (rank 3). Somatotropin, the growth hormone GH secreted by the pituitary gland, is central to several of the connections found by Swanson. Somatostatin is another hormone that inhibits the release of growth hormone and so is also an important aspect of the connections between SmC and Arginine. Other relevant terms include *body weight* ranked 4 under *Organism Attribute.* Again we see that P = 20 offers the best results.

Schizophrenia and Calcium-Independent Phospholipase A2

The starting point for this problem addressed by Smalheiser and Swanson (1998) is a report by Ross et al. (1997), which reported that levels of a calcium independent form of PLA2 are elevated in the serum of schizophrenic patients. The Ross et al. study established elevated PLA2 levels as a part of the characteristics of schizophrenia regardless of whether it is a cause or a consequence of the disease. It was also independently suggested in other reports that chronic oxidative stress may occur in schizophrenia. Smalheiser and Swanson found another study by Kua et al. (1995) working with rats intriguing in that they show that oxidative stress causes an elevation of PLA2 levels in lung, liver, and heart. The idea suggested by Smalheiser and Swanson was to combine the methods in the Ross et al. and Kua et al. studies to determine if oxidative stress also causes PLA2 levels in rat serum to become elevated. Such efforts might lead to the prediction that treating schizophrenia with antioxidants should reverse the elevation of serum PLA2.

We conducted a closed discovery process with schizophrenia as the A topic and calcium-independent PLA2 as

[9] PMID:7743205, published in January 1995; PMID:7528015, published in 1994.

[10] PMID:7827327, published in 1994.

[11] PMID:11080519, published in 2000.

[12] PMID:11259508, published in 2001.

TABLE 7. Closed discovery set sizes.

	Migraine			Raynaud's			Somatomedin			Indomethacin			Schizophrenia		
P	F	R	S	F	R	S	F	R	S	F	R	S	F	R	S
10	303	174	41	123	80	10	337	217	21	370	216	21	166	127	2
20	571	365	101	170	114	14	563	397	52	654	417	52	269	212	4
30	777	504	147	191	135	17	701	510	76	907	587	76	370	303	4
40	931	615	184	200	145	17	798	592	91	1131	750	98	417	346	4

the C topic limited to the pre-1998 human studies literature. We find the MeSH term *oxidative stress* ranked 3rd in the semantic type *Cell or Molecular Dysfunction* when P = 20 or 30 or 40. The term was not identified for P = 10. This MeSH term is the key connection between the two topics as identified by Smalheiser and Swanson. We also find *nerve degeneration* ranked 1 within the same semantic type for all settings of P. A search on PubMed reveals many articles discussing neurodegeneration in schizophrenia.[13] There is also some evidence indicating that enhanced PLA2 activity at least in rats may be related to neuronal degeneration.[14] This suggests that nerve degeneration could perhaps also have been identified as a connection between the A and C topics.

Discussion

Several summary observations may be made over the 2 open and 5 closed discovery problems addressed in this study. First, as expected the open problems are more challenging than the closed ones. However, our procedure was able to rank the key MeSH terms in the top 10 ranks within the appropriate semantic types for both open problems. We also observe that a more general version of the open procedure, one with a wider span ($N = 2$), degrades the ranks for both open problems. A different notion of generality is achieved by changing the semantic type selectors for the intermediate B terms. Interestingly, when using all 134 semantic types for B terms, the rank degrades for the Raynaud's–fish oils problem but it remains stable for the migraine-magnesium problem. Further research is required to better understand conditions responsible for these differences.

Our closed discovery algorithm has some very stable properties. Changing the value of parameter P beyond 20 does not significantly impact performance. Rankings for most key terms remain steady. In certain cases, higher values identify a few additional important terms. For example, in the migraine-magnesium problem, the hypoxia connection is only visible with P = 30 at least.

We now look at the number of ranked terms identified by each of our closed discovery processes. These are given in Table 7. Cell values indicate total number of terms in the set shown to the user after duplicate terms have been removed. Duplicate instances occur when a term has multiple semantic types. The column labels, *F, R,* and *S* represent varying conditions that will be defined shortly. But first we observe that, as expected, set size increases under all conditions with increasing *P*. We now focus our analysis on the row for P = 20. Columns labeled *F,* for full set, show set sizes when B terms in all 134 semantic types are displayed to the user. Size ranges from 170 to 654 (average: 445; average deviation: 181). Since these are fairly large sets from a user's perspective, ranking terms within semantic type is important. Columns labeled *R,* for reduced, depict set size after automatically removing 43 semantic types that seem obviously unrelated to these kinds of discovery problems. Examples of removed semantic types are: *Educational Activity, Health Care Related Organization,* and *Biomedical Occupation or Discipline.* For semantic types that are retained, we do not fine tune the terms contained using stop words, since these are likely to be problem specific. Under the reduced conditions, size ranges from 114 to 417 (average: 301; average deviation: 110). This represents a reduction of 21 to 36% (average: 31%) compared to the *F* sets. Columns labeled *S,* for select, represent set size when the semantic types are limited to those that actually contain the interesting terms. For example, in the Raynaud's problem, the 14 terms are from the union of *Cell Function, Physiologic Function,* and *Finding* terms (see Table 3). This column depicts a size range of 4 to 101 (average: 45; average deviation: 28). Compared to the *F* set, *S* represents a reduction of 82 to 99% (average: 91%). This analysis yields two observations. First, it is important to get input from the user on the semantic types of interest. Second, it is important to rank terms within semantic type. As shown before, most of our terms are ranked within the top 10 ranks of a semantic type.

We conclude that it is best to use a conservative open procedure with the minimal width of N = 1. For the closed procedure, we recommend P = 20 unless the user is better served with the smallest possible set of terms, in which case we recommend P = 10. These observations are made within the constraints of our experiments.

It is not easy to compare our results with those of previous replications since, as pointed out in the summary, our ranking strategies are different. We normalize term weights within each semantic type and these types define

[13] PMID:8637950, published in 1995; PMID:9278190, published in 1997.

[14] PMID:9436653, published in 1997.

independent term groups. In contrast, the other researchers offer a single ranking of all terms without any semantic groupings. Our user may select or eliminate semantic groups for further study with relative ease (possibly by looking at the top few terms in each group), whereas in the other approaches, the task of grouping the terms, if desired, is left to the user. Thus, terms belonging to more interesting semantic groups may be ranked below terms of less interesting groups. In general, our term rankings are consistently better than the previous results. However, we are unable to predict the ranks the others would have achieved had they also grouped their output by semantic type. There are other differences in our approaches. For example, in the open process given the user's preference regarding semantic types, we select B terms and build their profiles automatically. Given the same set of semantic type preferences, Weeber et al. (2000, 2001) manually select terms for searching. Similar manual decisions are made by Lindsay and Gordon (1996, 1999).

Finally, with reference to our original question, we see that text mining procedures built using only the MeSH metadata field of MEDLINE are more effective than the free-text-based methods explored by others. Moreover, our results have been obtained without the use of stop words, which in some of the other studies have included a few thousand phrases. The other advantage is that we do not have to deal with the challenge of correctly identifying phrases from free-texts and then correctly mapping them to MeSH or UMLS concepts. We have also not yet explored the fact that MeSH terms when applied to MEDLINE are classified into two groups, one more important than the other. The former group, called major MeSH terms, are marked with an asterisk. Also MeSH terms are often qualified with special phrases called subheadings. These provide further details on the particular aspect of the MeSH term addressed in the document. Major MeSH terms and subheadings will be the subject of future research. One disadvantage in our approach is that we will only be able to involve MEDLINE records in the text-mining process after they have been indexed with MeSH. Thus, a time lag is implied. Another aspect is that MeSH terms sometimes are more general compared to the actual concepts that appear in the free-text. Differences in granularity may become important in the long run and will also be addressed in future research.

Related Research

As stated in Blagosklonny and Pardee (2002), the era of conceptual biology is now upon us. The exponentially increasing amounts of published information and their complexity create obstacles to efficient research. At the same time, these vast resources offer an unparalleled opportunity to support hypothesis-driven, experimental research in biology. By automatically analysing published information and logically connecting concepts studied in seemingly unrelated fields, one can generate ideas for further research. Viewed in the reverse direction, one may also conduct preliminary explorations of tentative hypotheses by looking for supporting arguments in the published literature. Text mining applied to the domain of biomedicine is conceptual biology.

The research of Swanson and Smalheiser, Lindsay and Gordon, and Weeber et al., referred to thus far, form the basis of our research. More recently, Swanson et al. (2001) have explored the problem of categorizing viruses as biological weapons. Their approach is a natural extension of their previous research. Essentially, they use PubMed to explore four different properties of viruses that could be used as weapons. The properties pertain to virulence, airborne transmission, stability in air or aerosols, and transmission by agents such as insects. For each property, they conduct a PubMed search and then use ARROWSMITH to examine 3 property pairs (with virulence always one member of a pair). For each property pair, ARROWSMITH yields a list of virus MeSH headings that appear in both document sets. Statistical tests indicate that the 3 virus lists identify significant numbers of known virus weapons.

In new research, Weeber et al. (2003) use their open discovery procedure to look for novel therapeutic uses for thalidomide. A search on thalidomide initiates the process. The free-text of retrieved documents are mapped into UMLS concepts. The semantic type *Immunologic Factor* is used to select terms that are shown to a specialist, ranked by frequency. Selections made are used as B terms. Records retrieved for these B terms are analyzed in the same way and *Disease or Syndrome* terms extracted. After further filters, a set of 100 diseases are examined by the subject expert. Finally, four diseases are identified for which thalidomide may have a therapeutic role. Acute pancreatitis and chronic hepatitis C are two examples.

Other research exploring parallel text mining directions includes the many reports exploiting co-occurrence of concepts in the biomedical domain. Jenssen et al. (2001) generate a co-occurrence-based gene network called PubGene from MEDLINE for 13,712 named human genes. Each of PubGene's 139,756 links is weighted by the number of times the genes co-occur. Stapley and Benoit (2000) also exploit co-occurrence to generate a gene-gene map from MEDLINE documents containing the term *Saccharomyces cerevisiae.* While Stephens et al. (2001) also use co-occurrence data to postulate gene interactions, they go beyond simple frequency based counts. They also consider how frequently the gene term (or its synonyms) occurs in the document. Associations above a particular threshold are analyzed further using a list of relationship words such as *activates* and *cleaves.* Adamic et al. (2002) identify communities of genes. Starting with a co-occurrence-based gene network for a particular disease domain, communities are identified by repeatedly removing edges of highest betweeness (number of shortest paths traversing the edge). Applying this to the domain of colorectal cancer, they are able to identify interesting hypotheses linking genes that were, for

example, in the same community but had no edge between them.

Another approach aiming toward identifying the functional similarity between genes is that of Shatkay et al. (2000). The authors first identify for each gene a *kernel* document describing the gene's function. This document is then used to seek out similar documents from MEDLINE. Overlap in document sets are then used to estimate functional similarity between the source genes.

Association rules (Agrawal et al., 1993; Agrawal & Srikant, 1994; Feldman & Hirsh, 1997; Piatetsky-Shapiro & Frawley, 1991) are a dominant theme in text mining research (Blake & Pratt, 2001; Feldman & Dagan, 1995; Hristovski et al., 2001). These rules link pairs or larger groups of concepts and are assigned support and confidence values, scores that are commonly used in data-mining research. An association rule such as concept A → concept B indicates that there may be a potentially interesting directional association from A to B. Typically, these are discovered by exploiting the co-occurrence of A and B in the texts being mined. Hristovski et al. (2001) use association rules for literature-based knowledge discovery from MEDLINE. Although they do not exactly replicate any of the Swanson and Smalheiser discoveries, their overall approach derives from the open discovery approach. Given a pair of rules A → B and B → C, they apply the transitive operation to conclude that A → C. In their experiments, they use two time-based subsets of MEDLINE, 1990–1995 defining an old set and 1996–1999 defining a new set. Using association rules built from the old set, they were able to predict the majority of the novel A–C co-occurrences that were observed in the new set. Unfortunately, only a small percentage of the predicted relations were observed in the new set. For example, starting with Multiple Sclerosis as the A concept, they predict 79% of the novel term co-occurrences with this disease term in the new set but only 8% of their predictions were realized.

Conclusions

In this research, we have proposed and studied algorithms that focus on literature-based hypothesis generation that are designed within the discovery framework established by Swanson and Smalheiser. Our open and closed algorithms focus on the first dimension of the discovery processes, i.e., the identification and ranking of key terms. We have tested these algorithms on 2 open and 5 closed discoveries made by Smalheiser and Swanson. This is the most comprehensive replication study of their discovery problems to date. Where possible, our results were compared with those of other studies replicating the same discoveries. Comparisons indicate that our algorithms require far less manual intervention while displaying the key terms at very high ranks. Experiments were also conducted to explore the impact of the different parameters. These indicate that our algorithms are robust with generally predictable variations in performance over the parameter space. Our specific conclusions are:

1. For all discovery problems tested, almost all key terms are ranked within the top 10 positions under the appropriate semantic types. Thus, our term-weighting strategies effectively differentiate between MeSH terms within a semantic type. Also, after the user specifies the semantic types of interest, our algorithms successfully assign key terms to very high ranks. More generally, we conclude through this replication study that our MeSH profile–based discovery algorithms successfully discover key connections between topics. We have also been able to suggest new connections for several topic pairs.
2. In the open discovery process, the parameter N is best left at 1, which suggests that narrow searches for intermediate B terms are preferred.
3. In the open discovery process, it is important for problem-specific semantic types to be identified for selecting B terms. However, we temper this conclusion since our results on this aspect are not consistent. For example, using all semantic types yielded no degradation in performance for the migraine-magnesium problem. Further research is needed to understand these results.
4. In the closed discovery problems, the best results are achieved when parameter P is set to 20. Thus, when combining the two topic profiles, it is sufficient to consider only the top 20 ranked terms in each semantic type. Limiting our attention to the top 10 terms results in missing some key terms. Considering 20 or more terms yields no additional returns.
5. In the closed discovery process, by removing 43 of the 134 semantic types (since they were very general), we are able to trim the overall set of terms on average by 31% for the problems considered. However, if the user is able to specify the relevant semantic types to view, then the size of terms to consider drops on average by 91%.
6. Our study confirms that UMLS semantic types can be exploited in the discovery process and in addition may be used for organizing the results shown to the user. At present, during discovery we consider all semantic types corresponding to a MeSH term. Possibly, our algorithms could be made even more precise if instead we could automatically select the correct semantic type in the context of a given MEDLINE record. We will consider this "disambiguation" problem in the next phase of our research.

In future research, in addition to the specific points raised already, we plan explorations in several directions. First, encouraged by these good results, we will move toward exploring current discoveries in collaboration with biomedical scientists. In order to support this, we are presently creating a web-based front end for the user and involving database technology at the back end of the system. Second, we plan to explore methods by which one may assist a user with the second dimension of the discovery process, i.e., when the user has to peruse the documents to figure out the nature of the interaction underlying a suggested term rela-

tionship. Third, we will look at methods by which select free-text phrases may be used to augment the MeSH-based topic profiles. This may be a solution to the problem potentially caused by MeSH terms that are more general when compared to the terms in the title and abstract. Finally, we plan to identify and display semantic relationships between the different groups of terms shown to the user. These may be helpful to the user when analyzing output from the system.

Acknowledgments

The author gratefully acknowledges the comments on this research provided by Dr. Neil Smalheiser, University of Illinois at Chicago, and Dr. Marc Weeber, Erasmus University Rotterdam, The Netherlands. This research was partly accomplished while the author was a visiting faculty scholar at the National Library of Medicine, Bethesda, Maryland. She thanks the University of Iowa for the Faculty Scholar Award and ORISE for supporting her stay at NLM. This research is dedicated to the fine memory of the late Professor Jeffrey Katzer of Syracuse University.

References

Adamic, L.A., Wilkinson, D., Huberman, B.A., & Adar, L. (2002). A literature based method for identifying gene-disease connections. In Proceedings of the IEEE Computer Society Bioinformatics Conference, Stanford, CA.

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. In Proceedings of VLDB, International Conference on Very Large Data Bases, 487–499.

Agrawal, R., Imielinksi, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of ACM SIGMOD, International Conference on Management of Data, Washington, DC, May 26–28, 1993, 207–216.

Andrade, A., & Valencia, A. (1998). Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. Bioinformatics, 14(7), 600–607.

Aronson, A. & Rindflesch, T.C. (1997). Query expansion using the UMLS Metathesaurus. In Proceedings of AMIA, Annual American Medical Informatics Association Conference, Nashville, TN, October 25–29, 1997, 485–489.

Blagosklonny, M.V., & Pardee, A.B. (2002). Unearthing the gems. Nature, 416, 373.

Blake, C. & Pratt, W. (2001) Better rules, fewer features: A semantic approach to selecting features from text. In Proceedings of ICDM, IEEE International Conference on Data Mining, San Jose, CA, November 29–December 2, 2001, 59–66.

Chaussabel, D. & Sher, A. (2002). Mining microarray expression data by literature profiling. Genome Biology, 3(10):research0055.1–0055.16.

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, 8, 158–164.

Fayyad, U.M. & Uthurusamy, R. (1996). Data mining and knowledge discovery in databases (Introduction to the special section). Communications of the ACM, 39(11), 24–26.

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases. In Proceedings of KDD, International Conference in Knowledge Discovery and Data Mining, Montreal, Canada, August 20–21, 1995, 112–117.

Feldman, R. & Hirsh, H. (1997). Exploiting background information in knowledge discovery from text. Journal of Intelligent Information Systems, 9(1), 83–97.

Feldman, R., Aumann, Y., Amir, A., Klosgen, W., & Zilberstien, A. (1997). Maximal association rules: A new tool for mining for keyword co-occurrences in document collections. In Proceedings of KDD, International Conference in Knowledge Discovery and Data Mining, Newport Beach, CA, August 14–17, 1997, 167–170.

Gallai, V., Sarchielli, P., Coata, G., Firenze, C., Morucci, P., & Abbritti, G. (1992). Magnesium levels in migraing. Results in a group of juvenile patients. Headache, 32(3), 32–35.

Gordon, M.D., & Lindsay, R.K. (1996). Toward discovery support systems: A replication, reexamination, 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.

Hahn, U., & Schnattinger, K. (1997). Deep knowledge discovery from natural language texts. In Proceedings of KDD, International Conference in Knowledge Discovery and Data Mining, Newport Beach, CA, August 14–17, 1997, 175–178.

Hearst M. (1999). Untangling text data mining. In Proceedings of ACL, Annual Meeting of the Association for Computational Linguistics (invited talk), University of Maryland, Maryland, June 20–26, 1999.

Hristovski, D., Stare, J., Peterlin, B., & Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in Medline and UMLS. In Proceedings of MedInfo Conference, London, England, September 2–5, 2001, 10(2), 1344–1348.

Jenssen, T.-K., Laegreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 28, 21–28.

Kua, C.-F., Cheng, S., & Burgess, J.R. (1995). Deficiency of vitamin E and selenium enhances calcium-independent phospholipase A2 activity in rat lung and liver. Journal of Nutrition, 125, 1419–1429.

Lent, B., Agrawal, R., & Srikant, R. (1997). Discovering trends in text databases. In Proceedings of KDD, International Conference on Knowledge Discovery, NewPort Beach, CA, August 14–17, 1997, 227–230.

Lindsay, R.K., & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50(7), 574–587.

Masys, D.R., Welsh, J.B., Fink, J.L., Gribskov, M., Klacansky, I., & Corbeil, J. (2001). Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics, 17(4), 319–326.

Piatetsky-Shapiro, G., & Frawley, W.J.E. (1991). Knowledge discovery in databases. Cambridge, MA: MIT Press.

Ross, B.M., Hudson, C., Erlich, J., Warsh, J.J., & Kish S.J. (1997). Increased phospholipid breakdown in schizophrenia: Evidence for the involvement of a calcium independent phospholipase A2. Archives of General Psychiatry, 54, 487–494.

Shatkay, H., Edwards, S., Wilbur, W.J., & Boguski, M. (2000). Genes, themes and microarrays. Using information retrieval for large-scale gene analysis. In Proceedings of Intelligent Systems for Molecular Biology, La Jolla, California, 317–328.

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. (1998). Calcium-independent phospholipase A2 and Schizophrenia. Archives of General Psychiatry 55(8), 752–753.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28, 111–121.

Srinivasan, P. (2001). MeSHmap: A text mining tool for MEDLINE. In Proceedings of AMIA, the Annual Conference of the American Medical Informatics Association, Washington, DC, 642–646.

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

Stapley, B.J., & Benoit, G. (2000). Bibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In Proceedings of PSB, Pacific Symposium on Biocomputing, Hawaii, January 4–9, 2000, 5, 526–537.

Stephens, M., Palakal, M., Mukhopadhaya, S., Raje, R., & Mostafa, J. (2001). Detecting gene relations from MEDLINE abstracts. In Proceedings of PSB, Pacific Symposium on Biocomputing, Hawaii, January 3–7, 2001, 483–496.

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

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

Swanson, D.R. (1990). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine, 33(2), 157–179.

Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures. Artificial Intelligence, 91, 183–203.

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

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., Jong-van den Berg, L., & Vos, R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. In Proceedings of AMIA, the Annual Conference of the American Medical Informatics Association, November 4–8, 2000, 903–907.

Weeber, M., Klein, H., Berg, L., & Vos, R. (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, 52(7), 548–557.

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