College of Information Science and Technology

Mining Hidden Connections among Biomedical Concepts from Disjoint Biomedical Literature Sets through Semantic-Based Association Rule

Xiaohua Hu[1], Xiaodan Zhang[1], Guangren Li[2], Illhoi Yoo[1], Xiaohua Zhou[1], Xuheng Xu[1], Daniel Wu[1]

[1]College of Information Science and Technology, Drexel University, Philadelphia, PA 19104
thu@cis.drexel.edu
[2]Faculty of economy, Hunan University, Changsha, China

Abstract. The novel connection between Raynaud disease and fish oils was uncovered from two disjointed biomedical literature sets by Swanson in 1986. Since then, there have been many approaches to uncover novel connections by mining the biomedical literature. One of the popular approaches is to adapt the Association Rule (AR) method to automatically identify implicit novel connections between concept A and concept C from two disjointed sets of documents through intermediate B concept. Since A and C concepts do not occur together in the same data set, the mining goal is to find novel connection among A and Cconcepts in the disjoint data sets. It first applies association rule to the two disjointed biomedical literature sets separately to generate two rule sets (A → B, B → C), and then applies transitive law to get the novel connections A→C. However, this approach generates a huge number of possible connections among the millions of biomedical concepts and a lot of these hypothetical connections are spurious, useless and/or biologically meaningless. Thus it is essential to develop new approach to generate highly likely novel and biologically relevant connections among the biomedical concepts. This paper presents a Biomedical Semantic-based Association Rule System (Bio-SARS) that significantly reduce spurious/useless/biologically irrelevant connections through semantic filtering. Compared to other approaches such as LSI and traditional association rule-based approach, our approach generates much fewer rules and a lot of these rules represent relevant connections among biological concepts.

1 Introduction

The problem of mining novel hidden connections among biomedical concepts from biomedical literature was exemplified by Swanson's pioneering work on Raynaud disease/fish-oil discovery in 1986 [2]. According to Swanson [2] [5], connections

among biomedical concepts can be public, yet undiscovered, if independently created fragments of knowledge and information are logically related but never retrieved, interpreted and study together. In other words, two complementary and non-interactive literature sets of articles (independently created fragments of knowledge), when they are considered together, can reveal useful information of scientific interest not apparent in either of the two sets alone [2] [5].

Swanson formalizes the procedure to discover hidden connections from biomedical literatures as follows: Consider two separate literature sets, CL and AL, where the documents in CL discuss concept C and documents in AL discuss concept A. Both of these two literature sets discuss their relationship with some intermediate concepts B (also called bridge concepts). However, their possible connection via the concepts B is not discussed together in any of these two literature sets as shown in Figure 1. Simply, Swanson's model (aka ABC model) can be described as the process to induce "A implies C", which is derived from both "A implies B" and "B implies C"; the derived knowledge or relationship "A implies C" is not conclusive but hypothetical. For example, Swanson tried to uncover novel suggestions for what (B) causes Raynaud disease (C) or what (B) are the symptoms of the disease, and what (A) might treat the disease as shown in Figure 1. Through analyzing the document set that discusses Raynaud disease he found that Raynaud disease (C) is a peripheral circulatory disorder aggravated by high platelet aggregation (B), high blood viscosity (B) and vasoconstriction (B). Then he searched these three concepts (B) against Medline to collect a document set relevant to them. With the analysis on the document set he found out those articles show the ingestion of fish oils (A) can reduce these phenomena (B); however, no single article from both document sets mentions Raynaud disease (C) and fish oils (A) together. Putting these two separate literatures together, Swanson hypothesized that fish oils (A) may be beneficial to people suffering from Raynaud disease (C). This hypothesis that Raynaud disease might be treated by fish oil was hidden in the biomedical literature until Swanson uncovered it by using literature-based discovery. This novel hypothesis was later clinically confirmed by DiGiacomo in 1989[3].Later on, Swanson used the same approach to uncover 11 connections of migraine and magnesium [10] One of the drawbacks of Swanson's method is that the method requires large amount of manual intervention and very strong domain knowledge, especially in the process of qualifying the intermediate concepts Swanson call the "B" concepts. In order to reduce the dependence of domain knowledge and human intervention and to automate the whole process as much as possible, several approaches [6][8][9][11] have been developed to automate this discovery process based on Swanson's method. They have successfully not only replicated the Raynaud disease/fish-oil and migraine/magnesium discovery but also discovered new treatments for other diseases such as thalidomide [11]. Even though these research works have produced valuable insights into new hypothesis, however, substantial manual intervention has been required to reduce the number of possible connections. Specially, for association rule approaches [1] [8], they all did not utilize semantic information to automatically reduce the huge number of possible connections among the biomedical concepts.. This will be very time consuming and produce a lot of spurious/meaningless hypothesis. In this paper, we present a fully automated approach for mining hidden connections from biomedical literature. Our approach replaces manual ad-hoc pruning by using semantic knowledge from

biomedical ontologies. We apply semantic knowledge to association rule mining technique to discover novel connections between concepts. Unlike other approaches, our method utilizes both association rule technique and biomedical ontologies to automatic discover semantically related but implicit novel connections between concepts. When a new rule A→B is generated, our algorithm will automatically check the semantic types of both concepts. If their semantic type is unrelated, this rule will be filtered out. We use semantic information to manage and filter the sizable branching factor in the potential connections among a huge number of medical concepts. In order to solve the ambiguity problem of the biomedical terms and to discover novel hypotheses from a huge search space of possible connections among the biomedical concepts in an effective and efficient way, we utilize the biomedical ontologies, such as UMLS and MeSH. Our semantic-based association rule mining algorithm utilizes semantic knowledge (e.g., semantic types, semantic relations and semantic hierarchy) on the bridge concepts and the target concepts to filter out those irrelevant association rules and thereby meaningless connections between the concepts. The details are described in Section 3. The rest of the paper is organized as follows. Section 2 briefly discusses the relevant works that have improved Swanson's model. Section 3 describes our method in detail. The experimental results are presented in Section 4. Conclusion and future direction are discussed in Section 5.

2 Related Work

Several algorithms have been developed to overcome the limitations of Swanson's approach. Hristovski, et al. [4] used the MeSH descriptors rather than the title words of the documents. They use association rule algorithms to find the co-occurrence of the words. Their methods find all *B* concepts as bridges that are related to the starting concept C. Then all *A* concepts related to *B* concepts are found through Medline searching. But in Medline each concept can be associated with many other concepts, the possible number of B→C and A→B combinations can be extremely large. In order to deal with this combinatorial problem, the algorithm incorporates filtering and ordering capabilities [7] [8] [9]. Pratt and Yetisgen-Yildiz [8] used Unified Medical Language System (UMLS) concepts instead of MeSH terms assigned to Medline documents. Similar to Swanson's method, their search space is limited by only the titles of documents for the starting concept. They can reduce the number of terms (*B* concepts and *A* concepts) by limiting the search space. In addition to that, they reduce the number of terms/concepts by pruning out terms that are "too general" (e.g., terms such as problem, test, etc.), "too closely related to the starting concept", and "meaningless". They defined a term "too general" if the term is found in the titles of more than 10,000 documents. For "too closely related to the starting concept", they tracked all the parents and children concepts of the starting concept and then eliminated the related terms. To avoid "meaningless" terms, they followed the same method as in [4], manually selected a subset of semantic types to which the collected terms should belong. Before generating association rules, they tried to group the concepts (*B* or *A* concepts) to get a much coarser level of synonyms. Then, they removed "too general" concepts by looking at their UMLS hierarchy level and

non-UMLS concepts. With the qualified and grouped UMLS concepts, they used the well-known Apriori algorithm [1] to find correlations among the concepts. Although they managed to simulate Swanson's migraine-magnesium case only through concept grouping, their method still requires strong domain knowledge, especially on selecting semantic types for *A* and *B* concepts and also some vague parameters on defining "too general" concepts. Also above approach treats Apriori algorithm as an independent process and does not integrate semantic information in the frequent item sets generating process. Srinivasan [9] viewed Swanson's method as two dimensions. The first dimension is about identifying relevant concepts for a given concept. The second dimension is about exploring the specific relationships between concepts. However, only Srinivasan [9] deals with the first dimension. The key point of this approach is that MeSH terms are grouped into the semantic types of UMLS to which they belong. However, only a small number (8 out of 134) of semantic types are considered since the author believes those semantic types are relevant to *B* and *A* concepts. For each semantic type, MeSH terms that belong to the semantic type are ranked based on the modified TF*IDF. There are some limitations in their method. First, the author used manually-generated semantic types for filtering. Second, the author applied the same semantic types to both *A* and *B* terms. Because the roles of *A* and *B* terms for *C* term are different, different semantic types should be applied.

These research works have made significant progress on Swanson's method. However, none of the approaches considers the specific semantic relationships. For association rule approaches [1] [8], none of them apply semantic knowledge directly to the rule generating process. The association problem should be tackled by not only the information measure but also the semantic information among the concepts. In contrast, we focus on developing fully automated approaches to this problem based on the semantic knowledge about the medical concepts and their relationships. We use semantic information to automatically filter out irrelevant A→B and B→C rules and thus prune semantically unrelated medical concepts and bogus or non-interesting relationships among the medical concepts. The details are discussed in Section 3.

2.1 Semantic-based Mining Model for Novel Connections

We introduce a semantic-based mining model that explains how relationships or associations among concepts can be semantically induced. The model relies on biomedical anthologies, such as UMLS [http://umlsks.nlm.nih.gov] and MeSH [http://www.nlm.nih.gov/MeSH/MeSHhome.html] for identifying biomedical concepts and their semantic types and semantic relationships among them. It is based on the use of the semantic network in UMLS to identify meaningful correlations among concepts. Thus, we first briefly introduce MeSH and UMLS we use as biomedical anthologies and then we propose the semantic-based mining model for novel connections including the algorithm.

2.2. Biomedical Ontology

MeSH. The main purpose of Medical Subject Headings (MeSH) is to index Medline articles using the controlled vocabulary ("Descriptors" in NLM's term) and the thesaurus ("Entry terms" in NLM's term); thus MeSH can be used for cataloging the articles. During the process of indexing articles (after reading full versions), MeSH concepts are assigned to each Medline article. When MeSH terms are assigned to Medline documents, around 3-5 MeSH terms are set to "MajorTopic" which represents the document very well. We use MeSH terms assigned to the Medline documents since we believe that MeSH terms (especially MeSH descriptors, assigned as "Major Topic") represent documents more precisely.

UMLS. Medical Language System (UMLS) provides a mechanism for integrating all the major biomedical vocabularies including MeSH. UMLS consists of three knowledge sources; Metathesaurus, Semantic Network, and SPECIALIST lexicon. Metathesaurus as a core is organized by concepts (meaning), synonymous terms are clustered together to form a concept, and concepts are linked to other concepts by means of various types of relationships to provide the various synonyms of concepts and to identify useful relationships between different concepts [7]. All concepts are assigned to at least one semantic type as a category. For example, the term *Raynaud Disease* has a semantic type [*Disease or Syndrome*], and *Fish Oils* has a semantic type [*Biologically Active Substance*]. Currently, there are 135 semantic types. Each semantic type has at least one relationship with other semantic types. At this time of writing, there are 54 relations. Both the semantic types and semantic relationships are hierarchically organized.

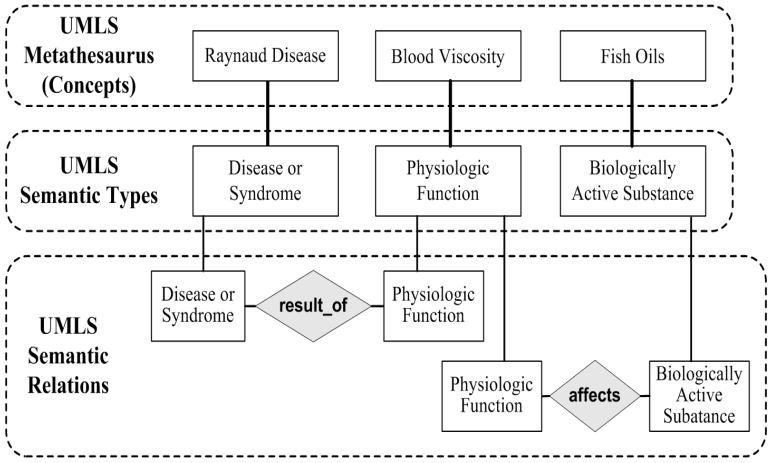

Figure 1. An illustrative example of the UMLS

Since most MeSH terms from Medline documents, are included into UMLS Metathesaurus Concepts, we know the semantic types of MeSH terms. Thus, given

two MeSH terms, we can derive the relationship between them from their semantic relation. Figure 2 shows the relationships of concepts, semantic types, and semantic relations of Raynaud Disease, Blood Viscosity and Fish Oils.

3 The Algorithm Bio-SARS.

We have developed a semantic association rule based literature mining system, called *Biomedical Semantic-based Association Rule System (Bio-SARS) as shown in Figure 2*. The input is a Medline search keyword as a "MajorTopic" MeSH term plus date range, the possible semantic relationships between *C* (the starting concept) and B concepts. Our algorithm takes the full advantage of the semantic knowledge in UMLS to check the semantic types for *B* to find out whether they have direct connections with that of concept C. And thus we only discover those relevant B→C rules. Accordingly, we use this method to discover only those A→B rules with this semantic relations. For example, for the 529 documents on "Raynaud disease" from 1980 to 1985, we extract 2746 medical terms after applying stop word list (including some of the most frequent used 325 MeSH terms) check and UMLs medical term check . However, if we apply semantic check to these terms, we will only extract 2036 semantic related terms, which will filter a lot of unrelated B concepts and thus help save computing time of association rule algorithm.

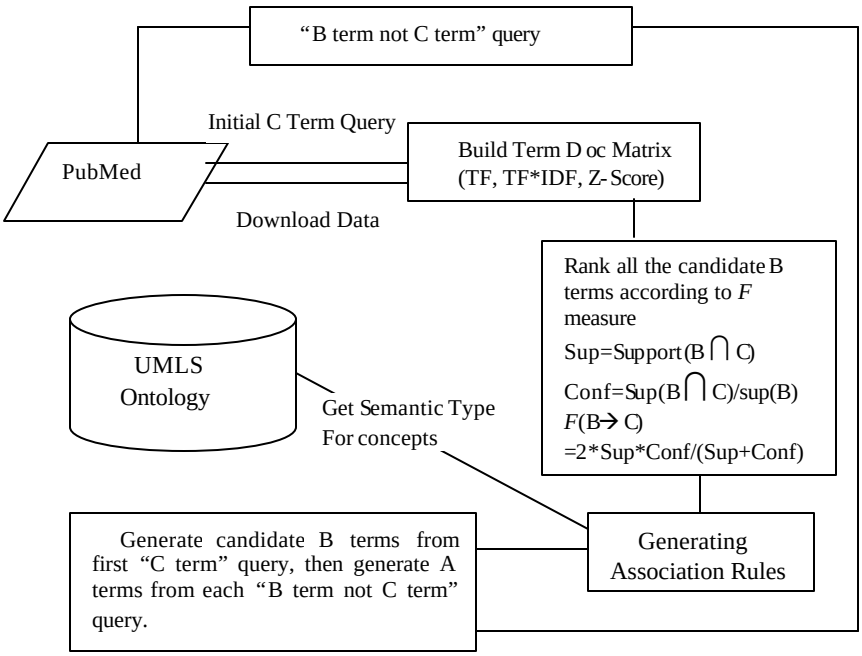

Figure 2. Biomedical Semantic-based Association Rule System (Bio-SARS). Support(B) means that the possibility that B occurs. Accordingly, Support (B∩ C) means the possibility that B and C occur together. Conf means confidence. Conf=Sup(B∩ C)/sup(B) means the confidence that B implies C (B→ C). *F* (B→ C) measure is a measure of confidence of B→ C. The larger the value is, the more we are confident that B→ C.

Procedural: Input starting concept *C* as MeSH term plus date range; Output: Target Concept List (*A* concepts)

Step1 Download the top k documents from PUbMED through query [concept C term + time period]

Step2 Extract all the terms from downloaded documents as candidate B terms (MeSH Heading, Title, and Abstract). Here we take all the terms that co-occur with C terms as stop word list for finding A terms that do not co-occur with C terms.

Step3 Compare the semantic types of C term with that of candidate B terms and then remove those B terms whose semantic types are not related to the semantic type of the C term

Step4 Build a matrix of terms by documents

Step5 Generate all B→ C association rules (B terms). Ranked them by $F = \frac{2\,Sup \times Conf}{Sup + Conf}$ measure. Select the top ***n*** B terms

Step6 For each B_i (i =1, 2, 3, …**n**) do

(1) Download the top k documents from PUBMED through query [***B Not C term***+ time period]. The time period is the same as Step1.

(2) Extract all the terms as candidate A terms from downloaded documents (MeSH Heading, Title, Abstract)

(3) Remove all those A terms that co-occur with concept C term.

(4) Compare the semantic types of Bi term and the entire candidate A terms and then remove those A terms whose semantic types are not related to the semantic type of Bi

(5) Build a matrix of terms by documents

Generate all A→ B association rules (A terms). Ranked them by $F = \frac{2\,Sup \times Conf}{Sup + Conf}$ measure. Select the top n A terms

Step7 List all A → B rules (A terms)

Below we explain each step in great details using the Raynaud disease as our example.

Step 1. Bio-SARS first download 529 documents in XML format from PubMed through query "Raynaud disease" [major] 1980:1985 [edat]. "[major]" indicates "major topic", while "[edat]" indicates "publication date".

Step 2. Extract all the terms as candidate B terms from document fields including MeSH headinglist, Title and Abstract. Here we take all the terms that co-occur with C terms as stop word list for finding A terms that do not co-occur with C terms. For generating candidate B terms, a stop word list, part of speech tagging, and UMLS medical term validation check are applied. From this step, we generated 570 B candidate terms. Below is part of term doc matrix we generated from the field called MeSHHeadingLists of each document. For association rule processing, we use the mean as threshold of cell values of each row vector that represents term vector. For cell value over threshold, we set it to 1, otherwise to 0.

Table 1. Example of term doc matrix

Terms	**Doc1**	**Doc2**	**Doc3**
Bleomycin	1	0	0
Lymphoma	1	0	0
sarcoma, Kaposi	1	0	0
blood pressure	0	1	0
Capillaries	0	1	0
Cyanosis	0	1	0
regional blood flow	0	1	0
vascular diseases	0	1	0
arterial occlusive diseases	0	0	1
ganglia, sympathetic	0	0	1
Hyperhidrosis	0	0	1
Ischemia	0	0	1
Sympathectomy	0	0	1
lupus erythematosus, systemic	0	1	0
adrenergic beta-antagonists	0	0	1

Step 3. Compare the semantic types of concept C term with that of all the candidate B terms to check whether they have direct semantic relations. If they don't have semantic relations, the candidate B terms will be removed from the B term list. For example, the last two terms in the table above were filtered out.

Step 4. There have 418 B terms extracted from the last step, the system will build a 418*300 term by documents matrix.

Step 5. Generate all the B→C rules are generated from this matrix, rank these rules according to $F = \frac{2\,Sup \times Conf}{Sup + Conf}$ to measure the closeness between concept B term and "raynaud disease". We calculate all the F(B→"Raynaud disease") value. At last, we get a ranked B term list. "Sup" indicates support of rule B→C, while "Conf" indicates confidence of rule B→C.

Table 2 Example of B→ C rules and F value (Since "Raynaud disease" is the initial term for all these B terms, we would not put the rule as B→C format and just put B terms in the table). Signal "!!" is used to separate two terms. These two terms both implies C term (B1&B2 →C). Here B term is at last allowed to be two terms.

Rules	***F* value**	**Rules**	***F* value**
occupational diseases	0.214	esophageal diseases	0.089
Plethysmography	0.170	ischemia	0.083
Calcinosis	0.142	epoprostenol	0.083
blood pressure	0.130	random allocation	0.083
regional blood flow	0.125	arterial occlusive diseases	0.070
Nifedipine	0.125	centromere	0.070
Telangiectasis	0.119	vasoconstriction	0.070
calcinosis!!telangiectasis	0.107	calcinosis!!esophageal diseases	0.070
Capillaries	0.101	blood viscosity	0.064
connective tissue diseases	0.095	blood flow velocity	0.064

Step 6. The system uses the top ranked 100 B terms to discover new concept. For each B concept, the system will build a term by document matrix to discover candidate A terms. For example, if the B term happens to be "blood viscosity", then the system will submit query —"blood viscosity" not "raynaud disease" 1980:1985 [edat] to PubMed and download 300 documents by default, which will guarantee B and C does not co-occur each other in the same document and thus reduce the possibility that the candidate A terms extracted co-occur with C term. Then, the system will extract all the terms from these downloaded documents as candidate A terms. For each document, the extracted terms will be from fields such as MeSH headinglist, Title and Abstract. A stop word list, part of speech tagging, and

UMLS medical term validation check will be applied. Then, the system will remove all the terms that co-occur with "Raynaud disease". Next, the system will build a matrix from these candidate A terms by their according documents. Last, we calculate all $F(A \rightarrow B)$. Below are some well ranked A terms that imply B term "platelet aggregation". All these terms do not co-occur with C term "Raynaud disease". So these are all novel connections. Although there are many other novel connections, as for space, we only list some highly ranked connections.

Table 3 Example A→B (platelet aggregation) rules and F value

Rules	*F* value	Rules	*F* value
adenosine diphosphate	0.358	Indomethacin	0.076
Epinephrine	0.197	Tiaramide	0.006
Glycoproteins	0.119	cyclic nucleotide phosphodiesterase	0.006
platelet membrane glycoproteins	0.119	Tiaramide	0.006
platelet activating factor	0.118	Anagrelide	0.006
Ristocetin	0.113	Esculetin	0.006
cyclic AMP	0.101		

Step 7. Last, we get a ranked A term list. We selected top 13 highly ranked A terms such as adenosine diphosphate, epinephrine glycoproteins, platelet membrane glycoproteins, ristocetin, etc.

4 Experimental Results

In our experiments, we reimplemented and evaluated the two existing approaches: latent Semantic Indexing (LSI)-based [6, 11, 12, 13] and standard association-rule based [1, 8] (AR) for mining the hidden links and compared them with our Bio-SARS on two of Swanson's famous medical discoveries, "*Raynaud Disease – Fish Oils*" and "*Migraine – Magnesium*". Figure2 shows how our prototype system works.

The desire of reimplementation and evaluation is from the fact that Latent Semantic Indexing (LSI) and Association Rule (AR) are two potential tools for discovering implicit knowledge [1,6,8,11,12,13]. Gordon [13] used LSI on Swanson's Raynaud's –Fish oils discovery and showed that LSI might be a useful technique in literature-based discovery: during the search for intermediate literatures, it fairly closely reproduces (but extends) the same set of highly ranked terms and phrases that Gordon and Lindsay [14] have shown are a useful starting point literature-based discover; in helping identify potential discovery literatures, LSI can be used by factoring a set of documents with a suspected intermediate literature, or by analyzing the larger literature that forms the universe of discourse. Association rule mining technique [8] helps identify correlations among concepts, and uses those correlations for open-ended discovery and identified a large set of intermediate terms between

magnesium to *migraine*.

In our experiments, we will explore the following questions for LSI, AR, SAR:

1. Will our method dramatically reduce unnecessary meaningless rules?
2. How do terms from different field (Mesh term, Title and Abstract) of a document from PubMed affect the experiment result?
3. How does cell value (TF, TFIDF, Z-Score) of a term document matrix affect experiment results?

4.1 LSI based Algorithm

For LSI, we specially use Singular Value Decomposition (SVD). SVD allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences [12]. For any matrix X (m by n), it can be decomposed as three matrices $T_0 S_0 D_0^T$ Here T (m by k) and D (k by n) are orthogonal matrices, while S (k by k) is a diagonal matrix. Each original term is now expressed as statistically independent factors [13] (represented by row of matrix $T_0 \times S_0$. The cosine between two row vectors reflects the extent to which two terms have a similar pattern. The larger the cosine is, the more similar the two terms are. Theoretically, this compare is better than standard cosine similarity. Thus, we use this technique to compare the similarities between the input term and all the other terms extracted from the documents. For example, for "Raynaud disease" as input C term, we can calculate the closeness of all the other terms after SVD analysis. Then we choose those terms that have good ranking as input B terms, thereby we can rank all the other terms that are disjointed with "Raynaud disease".

The procedure of LSI algorithm: Input: C term query; Output: Candidate A terms.

1. Download the top k documents from PUbMED through concept "C term" query within certain time period
2. Extract all the terms as B terms from MeSH headinglist, Title and Abstract after applying stop word list, part of speech tagging, and UMLS words validation check
3. Build a matrix of terms by documents and then analyze the matrix by SVD
4. Rank all the B terms according to the term vector cosine between concept C term and B terms and select top n B terms.
5. For each B_i (i =1, 2, 3, …n) do
   a) Download k documents from PUBMED through concept "B term" query within same time period
   b) Repeat step (2) to extract all the candidate A terms but remove all the terms co-occur with term C
   c) Repeat step (3)
6. Rank all the A terms according to the term cosine between A and C plus term cosine between B and C

4.2 Association rule based Algorithm (AR)

Association rules identify collections of data attributes that are statistically related in the underlying data. An association rule in our case is of the form B→A where B and A are disjoint conjunctions of attribute value pairs. For Rule: B→A, let $Sup = Support(B \cap A)$, $Conf = \frac{Support(B \cap A)}{Support(B)}$ we use $F = \frac{2Sup \times Conf}{Sup + Conf}$ to rank all the rules that satisfy B→A. For example, say, we want to calculate the *F* value of "blood viscosity"→ "fish oils". We first calculate the support of "blood viscosity" as B term, that is probability of the occurrence of "blood viscosity", as well as Support ("blood viscosity" n "fish oils"), From these two basic values, we accordingly calculate the confidence value of "B and A" and the *F* value of "blood viscosity"→ "fish oils". In this way, we rank all the B terms, and accordingly rank all the A terms. For B→A, since we always know B, better performance is guaranteed. For example, if we take "Raynaud disease" as input C concept, using method above, we can find out all the B concepts with good ranking. Then we use these B concepts as starting concept one by one to get all the A concepts. Here we take C concept as input, and then we calculate all B→C rules. Last we generate all A→B rules.

The procedure of AR algorithm: Input: C term query; Output: candidate A terms

1. Download the top k documents from PUbMED through concept "C term" query within certain time period.
2. Extract all the terms as candidate B terms from documents fields including MeSH headinglist, Title and Abstract after applying stop word list, part of speech tagging, and UMLS words validation check. We take all the terms that co-occur with C terms as stop word list for finding A terms that do not co-occur with C terms.
3. Build a matrix of terms by documents
4. Generate all B→C association rules (B terms) and rank all the B terms according to $F = \frac{2 Sup \times Conf}{Sup + Conf}$ and then choosetop n B terms.
5. For each B_i (i =1, 2, 3, …n) do
   a) Download k documents from PUBMED through "B Not C term" query within the same time period
   b) Remove terms that co-occur with C terms in "C term"
   c) Repeat step (2) to extract all candidate A terms
   d) Repeat step (3), (4) to build term doc matrix after removing unrelated terms
   e) Repeat step (5) to generate A→B rules (A terms)
6. List all A→B rules (A terms).

4.3 "Raynaud Disease – Fish Oils"

In our experiments, we use standard term cosine similarity ranking after LSI analysis to compare the term closeness. The initial query is "Raynaud Disease [major] edat:1980:1985". We download k=300 documents from PUBMED each time. We

have approximated the original term-document matrix using 100 (<300) orthogonal factors. We make six experiments all together. Each experiment has a different matrix according to the terms extracted only from MeSH HeadingList (MESH) or extracted both from [MeSH HeadingList and Title, Abstract] (MESHTAB), also according to the cell of matrix, term frequency (TF), term frequency and inverse document frequency (TFIDF), and Z-Score.

Table 4. LSI (Raynaud Disease—Fish Oil) In the table below, MeSH indicates that the terms are extracted only from MeSH heading field of each document, while MTAB means that terms are extracted from MeSH heading, Title, and Abstract field. TF, IDF and Z-Score are measures for the matrix cell value.

Selected Top B terms from which fish oil is discovered for each experiment	**B term is the # Closest term to Raynaud disease(C)**	**Fish oil (A) is the # closest term to B term**	**Term Document Matrix representation**
Plethysmography	17	766	MeSH TF
Arteriosclerosis	37	9253	
Eczema	41	1456	
Blood viscosity	70	300	
Plethysmography	17	483	MeSH TFIDF
Blood viscosity	70	2765	
Plethysmography	**17**	**475**	**MeSH ZScore**
Blood viscosity	70	442	
Arteriosclerosis	37	1693	MTAB TF
Eczema	41	1557	
Plethysmography	79	1466	
Eczema	53	2440	MTAB TFIDF
Arteriosclerosis	67	2097	
Eczema	52	1568	MTAB ZScore
Arteriosclerosis	67	1188	

In table 4, we only show those intermediate B terms from which fish oil is discovered. For example, for experiment MeSH+ZScore (bold character), we found a B concept plethvsmography ranked as 17 according to the distance to C concept "raynaud disease" and A concept "fish oil" is the 475 closest term to "plethvsmography". We also see that measures TF, TFIDF and ZScore don't affect

results too much, while adding title and abstract to MeSH terms does affect result. Plethysmography is an important B term since it occurs in four experiments. The reason that it does not come up with the other two experiments is that it ranks below 100 because we only find A terms close to the first 100 B terms. Besides, we also found some other B terms from which fish oil is discovered such as hypertension, arterial occlusive diseases, prostaglandins E, arteries, blood platelets, platelet aggregation, and collagen.

From the different ranking of term "fish oils", we can see that LSI might not be a good method for ABC discovering. Although better result might be achieved if we include some most frequently used MeSH terms in the stop list, it would not change the whole image of the ranking.

Table 5. Minimum # of B (intermediate terms) → C (Raynaud Disease) rules. MeSH term only means that terms are only extracted from MeSH heading field for each document. MeSH term+Title&Abstract means that terms are extracted from MeSH heading, Title and Abstract field.

The first B term that Generated fish oils (A term) in each of experiments with and without semantic type check	**Semantic Association Rule** / **Association Rule**	
	TF	TFIDF
Blood viscosity(B) (MESH term only)	418 / 570	418 / 570
platelet aggregation(B) (MESH term+Title&Abstract term)	1454 / 1952	1454 / 1952

Table 6. Minimum # of A (Fish oil) → B (intermediate terms) rules

The first B term that Generated fish oils (A term) in each of experiments with and without semantic type check	**Semantic Association Rule** / **Association Rule**	
	TF	TFIDF
Blood viscosity(B) (MESH term only)	18081 / 47888	16500 / 47888
platelet aggregation(B) (MESH term+Title&Abstract term)	124612 / 255750	66177 / 152461

In table 5&6, the term is the first B term from which term "fish oils" is discovered. Also TF means the cell value is set as term frequency, so does TFIDF. We use *F* measure to calculate the closeness between term pair such as A and B or B and C. From the experiment result in the tables, we can see that B term—Blood viscosity is recognized as the first B term to generate A term—fish oil in the experiments of MeSH+TF and MeSH+TF*IDF by both AR and SAR method, while "Platelet

aggregation" is for the experiments MeSHTAB+TF and MESHTAB+TFIDF by both AR and SAR method.

The minimum number of A→B and B→C rules is getting larger when adding terms from title and abstract to MeSH terms. TF and TFIDF do not affect results for the experiments using MeSH term only, however, they do affect for the experiments using both MeSH terms and Title & Abstract terms. This can be resulted from that all the MeSH term's TF is 1, while TF scales differently when adding title and abstract field.

Obviously, SAR reduces at least half of association rules whose semantic types don't match. For example, the minimum "fish oil"→"blood viscosity" rules for AR (Table 3) are 47888, while they are 18081 for SAR.

Table 7. Compare the ranking of top ranked A terms from which "fish oils" are discovered

A Term	Rank by AR	Rank by SAR
Blood viscosity	27	19
Blood platelets	32	23
Platelet Aggregation	33	24
Prostaglandins E	44	37

From table 7 we can see A term ranked by SAR all have better ranking than that ranked by AR. Fish oil is ranked 177 among 442 terms close to Blood viscosity for MESH+TFIDF experiment by SAR, while it's ranked 305 among 765 terms close to blood viscosity by AR.

4.4 "Migraine – Magnesium"

We conduct this experiment in the same way as the experiments on "Raynaud Disease". Here we choose time period between 1980 and 1984. In table 8, we only show those intermediate B terms from which magnesium is discovered. We also found similar result as "Raynaud disease—fish oils": magnesium does not have a good ranking. Besides the sample intermediate B terms in the table, we also found intermediate terms from which magnesium is discovered such as puerperal disorders, postpartum, hydrocortisone, ergotamine, gastrointestinal motility, phenethylamines, aerospace medicine, nicotinic acids, nimodipine, propranolol, blood platelets, cerebrovascular circulation, myotonia, chlorpromazine, chlorpromazine, iris, stress, and muscle contraction. These terms are all ranked within top 50 according to the term cosine value with C term migraine.

Table 8. LSI (Migraine – Magnesium). In the table below, MeSH indicates that the terms are extracted only from MeSH heading field of each document, while MTAB means that terms are extracted from MeSH heading, Title, and Abstract field. TF, IDF and Z-Score are measures for the matrix cell value.

Selected Top B terms from which magnesium	B term is the No. # Closest	Magnesium (A) is No. # closest	Term document

is discovered for each experiment	term to Migraine(C)	term to B term	matrix representation
Ergolines	9	259	
Nicergoline	10	256	MeSH
Benzamides	13	866	TF
Pre-eclampsia	14	332	
Ergolines	9	282	
Nicergoline	10	256	MeSH
Benzamides	13	822	TFIDF
Blood Pre-eclampsia	14	424	
Ergolines	9	739	
Nicergoline	10	256	MeSH
Benzamides	13	689	ZScore
Blood Pre-eclampsia	14	242	
Pre-eclampsia	8	535	MTAB
Benzamides	14	1320	TF
Pre-eclampsia	8	907	MTAB
Benzamides	14	1290	TFIDF
Pre-eclampsia	8	1019	MTAB
Benzamides	14	1517	ZScore

Table 9 Minimum # of B (intermediate term) →C (Migraine) rules. MeSH term only means that terms are only extracted from MeSH heading field for each document. MeSH term+Title&Abstract means that terms are extracted from MeSH heading, Title and Abstract field.

The first B term that Generated Magnesium (A term) in each experiment with and without semantic type check	Semantic Association Rule / Association Rule: TF	TFIDF
cerebrovascular circulation (B) (MESH term only)	474 / 674	474 / 674
ergotamine (B) (MESH term +Title&Abstract term)	1374 / 1895	1374 / 1895

Table 10 Minimum# of A(Magnesium)→B (intermediate B term) rules

The first B term that Generated Magnesium (A term) in each experiment with and without semantic type check	Semantic Association Rule / Association Rule: TF	TFIDF
blood platelets (B) (MESH term only)	4852 / 14670	4832 / 14623
ergotamine (B) (MESH term+Title&Abstract term)	25024 / 54410	17198 / 49218

In table 9 and 10, we also found SAR has dramatically reduced the number of B→C and A→B rules (terms) whose semantic types don't match. For example, experiment MESH TAB+TF with semantic type check generates fewer than half of A→B rules than that of MESHTAB+TF without semantic type check.

Besides term "Blood platelets" is ranked as 6, in experiment MESH+TF, we also found the following intermediate B terms such as "relaxation techniques" (7), "muscle contraction" (8), "food hypersensitivity" (12), serotonin (14), "ischemic attack, transient" (21), "calcium channel blockers" (25), "headache and relaxation techniques"(28) brain ischemia (34), aspirin (38), and spreading cortical depression (43), and vasodilation(44). All these terms have better rank in SAR experiment than in AR experiment.

From all above experiments, all the three questions are clearly answered. All these experiments indicate that Bio-SARS generates fewer novel but relevant connections than the standard association rule algorithm. Terms from different field (MESH, Title, and Abstract) do affect the experiment results, since ranking will change when more terms are added. Cell value basically does not affect LSI experiment, while it does affect AR and SAR experiment. The difference can be from that we setup threshold value for AR and SAR while we do not set up for LSI because threshold can easily skew the results of LSI method that reflects the whole semantic space very well and AR and SAR can only accept 1 or 0 value.

5 Conclusion and Future Work

This paper proposed a semantic based association rule mining method for Undiscovered Public Knowledge. For a given starting medical concept, it discovers new, potentially meaningful relations/connection with other concepts that have not been published in the medical literature before. The discovered relations/connections are novel and can be useful for domain expert to conduct new experiment, try new treatment etc.

The most significant novel feature of our SAR is that it dramatically reduces unnecessary meaningless semantic unrelated association rules to more quickly discover exact semantically related rules. Our method takes advantages of the biomedical ontologies, MeSH and UMLS and association rule text mining technique. There may be lots of ties of A concepts. The root cause of the problem is that the

relationships are assigned in the semantic type level instead of the concept level in UMLS. Because a semantic type contains lots of concepts and the relationships are assigned in the semantic level, the relationships among concepts are inevitably obscure, ambiguous or equivocal.

As our future research, we will reduce and rank A concepts in a semantic manner, which would be a challenging issue. For this problem, we may need more disease specialized biomedical ontology, such as Systematized Nomenclature of Medicine (SNOMED) [http://www.snomed.org/].

Acknowledgments. This research work is supported in part from the NSF Career grant (NSF IIS 0448023). NSF CCF 0514679 and The PA Dept of Health Tobacco Settlement Formula Grant (#240205, 240196,).

References

1. Agrawal, R., et al., Fast Discovery of Association Rules, *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, et al., Editors. 1995, AAAI/MIT Press.
2. Swanson, DR. Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7-18, 1986.
3. DiGiacome, R.A, Kremer, J.M. and Shah, D.M. Fish oil dietary supplementation is patients with Raynaud's phenomenon: *A* double-blind, controlled, prospective study, *American Journal of Medicine*, 8, 1989, 158-164.
4. Hristovski D, Stare J, Peterlin B, and Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. *Medinfo*. 2001, 10(Pt 2), 1344-8.
5. Swanson, DR. Undiscovered public knowledge. *Libr. Q*. 56(2):103-118, 1986.
6. Lindsay, R.K, and Gordon, M.D. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, 50(7):574-587, 1999.
7. National Library of Medicine (NLM), 2004AC UMLS Documentation, http://www.nlm.nih.gov/research/umls/documentation.html, 2004.
8. Pratt, Wanda and Yetisgen-Yildiz, Meliha, LitLinker: capturing connections across the biomedical literature, *K-CAP'03*, pp. 105-112, Sanibel Island, FL, Oct. 23-25, 2003
9. Srinivasan, P., Text mining: Generating hypotheses from MEDLINE, *Journal of the American Society for Information Science*, 2004, Vol. 55, No. 4, pp. 396-413
10. Swanson, DR. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526-557, 1988.
11. Weeber, M., Vos, R., Klein, H., de Jong-Van den Berg, L.T.W., Aronson, *A* & Molema, G. 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*,

12. Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K. Landauer and Richard Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 1990, vol. 41, no. 6, pp391-407

13. Michael D. Gordon, Susan Dumais, Using Latent Semantic Indexing for Literature Based Discovery Journal of the American Society for Information Science, 1998, vol. 49, no. 8, pp674-685

14. Gordon, M. D., & Lindsay, R. K. 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, 1996,47, pp116–128.