

A Semantic Approach for Mining Hidden Links from Complementary and Non-interactive Biomedical Literature*

Xiaohua Hu¹, Xiaodan Zhang¹, Illhoi Yoo¹, Yanqing Zhang²

¹College of Information Science and Technology, Drexel University,
Philadelphia, PA 19104

²Dept. of Computer Science, Georgia State University, Atlanta, GA 30302

{thu@cis, iy28@, xzhang@cis, xz37, daniel.wu}.drexel.edu, yzhang@cs.gsu.edu

Abstract

Two complementary and non-interactive literature sets of articles, when they are considered together, can reveal useful information of scientific interest not apparent in either of the two sets alone. Swanson called the existence of such hidden links as undiscovered public knowledge (UPK). The novel connection between Raynaud disease and fish oils was uncovered from complementary and non-interactive biomedical literature by Swanson in 1986. Since then, there have been many approaches to uncover UPK by mining the biomedical literature. These earlier works, however, required substantial manual intervention to reduce the number of possible connections. This paper proposes a semantic-based mining model for undiscovered public knowledge using the biomedical literature. Our method replaces manual ad-hoc pruning by using semantic knowledge from the biomedical ontologies. Using the semantic types and semantic relationships of the biomedical concepts, our prototype system can identify the relevant concepts collected from Medline and generate the novel hypothesis between these concepts. The system successfully replicates Swanson's two famous discoveries: Raynaud disease/fish oils and migraine/magnesium. Compared with previous approaches such as LSI-based and traditional association rule-based methods, our method generates much fewer but more relevant novel hypotheses, and requires much less human intervention in the discovery procedure

Keywords

Text Mining, Automatic Semantic Pruning, Biomedical Ontology, MeSH, UMLS, Swanson

1. Introduction

The problem of mining hidden links from complementary and non-interactive biomedical literature was exemplified by Swanson's pioneering work on Raynaud disease/fish-oil discovery in 1986 [11]. 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 [1] [12]. Swanson formalizes the procedure to discover UPK 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. 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

* This work is supported partially by the NSF Career grant IIS 0448023 and NSF CCF 0514679 and PA Dept of Health Tobacco Formula Grants

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 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 this paper, we present a fully automated approach for mining hidden links from biomedical literature. Our approach replaces manual ad-hoc pruning by using semantic knowledge from biomedical ontologies. We use semantic information to manage and filter the sizable branching factor in the potential connections among a huge number of medical concepts. Our method requires the minimum human intervention. Unlike other approaches [4] [8] [9], our method only requires the user to specify the possible semantic relationships between the starting concept and the to-be-discovered target concepts rather than possible semantic types of the target concepts and the bridge concepts. Our method utilizes semantic knowledge (e.g., semantic types, semantic relations and semantic hierarchy) on the bridge concepts and the target concept to filter out those irrelevant concepts and meaningless connections between the concepts.

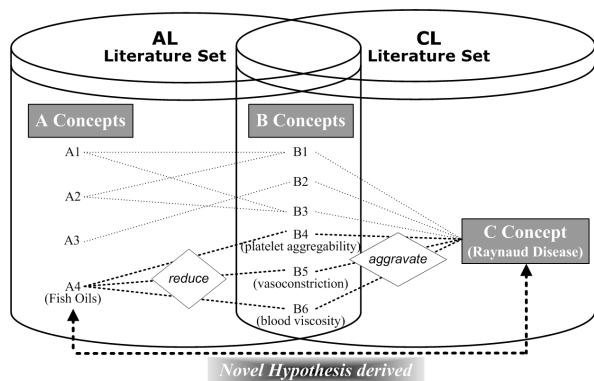


Figure 1. Swanson’s UPK model – the connection of Fish Oils and Raynaud Disease

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 $B \rightarrow C$ and $A \rightarrow B$ combinations can be extremely large. In order to deal with this combinatorial problem, the algorithm incorporates filtering and ordering capabilities [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 before generating association rules, they tried to group the concepts (B or A concepts) to get a much coarser level of synonyms. 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. 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. These research works have made significant progress on Swanson’s method. However, none of the approaches considers the specific semantic relationships. 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 prune irrelevant medical concepts and bogus or non-interesting relationships among the medical concepts. Our approach replaces manual ad-hoc pruning by using an existing biomedical ontologies. Our use of an intermediate set of automated identified semantic types helps to manage the sizable branching factor.

3. Semantic-based Mining Algorithm for UPK: Bio-SbKDS

We introduce a semantic-based mining algorithm Biological Semantic-based Knowledge Discovery System (Bio-SbKDS) to discover the hidden relationships or associations among biomedical concepts. The algorithm uses the semantic types and semantic relationships from the ontology UMLS.

Unified 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. 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. Figure 2 shows the relationships of concepts, semantic types, and semantic relations of *Raynaud Disease*, *Blood Viscosity* and *Fish Oils*.

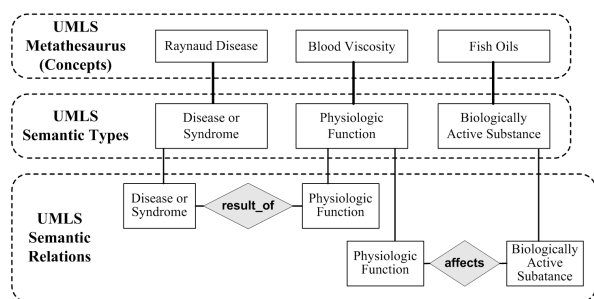


Figure 2. An illustrative example of the UMLS

Our algorithm takes a full advantage of the semantic knowledge in UMLS to select appropriate semantic types for *B* and *A* concepts through mutual qualifications and to identify relevant *B* and *A* concepts. The advantage of the algorithm is that, using only initial relations (possible relationships between *C* concept and *A* concepts), all the semantic types for both *B* concepts and *A* concepts are automatically derived using the biomedical ontology (UMLS). Because there must be at least one relationship between the semantic types for *B* and the semantic types for *A* concepts, the derived semantic types for *A* and *B* concepts are mutually qualified by considering their relationships. In Bio-SbKDS, 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 the to-be-discovered target concepts, and the role of the keyword for the initial semantic relations. For example, if the starting concept is *Raynaud Disease*, the relations selected are “treats” and “prevents” because we try to find something (the

target concepts *A*) that “treats” or “prevents” *Raynaud Disease*.

Algorithm Bio-SbKDS

Input: Starting concept *C* as MeSH term plus date range, the initial semantic relations ISR between the starting concept and the to-be-discovered target concept, the role of keyword for possible relations (subject or object)

Output: Target Concept List (*A* concepts)

Step 1: Find the semantic types *ST_C* of the starting concept *C* from the ontology UMLS;

Step 2: Find all the possible semantic types of the to-be-discovered concepts *B* related to *ST_C*; the semantic types derived are called *ST_B_can* (can means candidates), and are used as the category restriction for *B* concepts.

Step 3: Extract all semantic types related to ISR, which are the candidate semantic types for the to-be-discovered target concepts *A*, the result is denoted as *ST_A_can*.

Step 4: Extend *ST_A_can* obtained in STEP 3 by following through the ISA relations; the extended semantic types are called *ST_A_can_ext*.

Step 5: Check if there are relations between *ST_B_can* and *ST_A_can_ext* and also if the two semantic type sets pass the relation filter. If not, such semantic types are dropped from their semantic type list. After removing irrelevant semantic types, *ST_B_can* becomes *ST_B* and *ST_A_can_ext* becomes *ST_A*.

Step 6: Search the biomedical literature to get all the documents *CL* related to *C*; *CL* is the source of *B* concepts. Then, extract MeSH terms from *CL*; the terms are called *B_{can}*.

Step 7: Apply *B* concept category restriction (*ST_B*) to *B*; selecting the terms that only belong to at least one semantic type of *ST_B*. In addition, Bi-Decision Maker [5] further qualifies *B_{can}*. Here, the top ranked *B* terms, called *B_{top}*, are selected.

Step 8: Search all *B_{top}* terms to get all the documents *AL*; *AL* is the source of the to-be discovered *A* concepts. Then, extract MeSH terms from *AL*; the terms are called *A_{can}*.

Step 9: Apply *A* concept category restriction (*ST_A*) to *A_{can}*. In addition, Bi-Decision Maker further qualifies *A_{can}*.

Step 10: From *A_{can}*, retain those not co-occurred with *C* concept in Medline. The top ranked *A* concepts are selected.

Figure 3 shows the data flow of the procedure of mining the hidden links. Each number circled in Figure 3

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

Step 1: The semantic type of the starting concept C (ST_C) is identified through UMLS semantic network. At this time, only a MeSH term is allowed as a starting concept because the semantic type of the starting concept is used to construct the semantic type list for the B terms. For example, for the *Raynaud disease*, its semantic type is [*Disease or Syndrome*].

Step 2: All the semantic types (ST_{B_can}), which have at least one of the relations in the relation filters with ST_C (the semantic type of the keyword), are selected by considering the role of the initial keyword (i.e. as subject or as object). For example, in Table 1 [*Physiologic Function*] and [*Steroid*] are selected because the role of the initial keyword is set as an object on the interactive system and the relation filter

includes “*process_of*”, “*result_of*”, and “*causes*”; just regarding each record in Table 1 as a sentence (e.g. Steroid causes Disease or Syndrome). The filter relations between C and B are *process_of*, *result_of*, *manifestation_of*, and *causes*. The semantic types collected (ST_{B_can}) are used for the semantic types of B terms as category restriction. This is based on the fact that B terms have at least one relationship with C term.

Table 1. Semantic relations for some semantic types

Semantic Types (as subjects)	Relation	Semantic Types (as objects)
Physiologic Function	<i>process_of</i>	Disease or Syndrome
Physiologic Function	<i>result_of</i>	Disease or Syndrome
Steroid	<i>causes</i>	Disease or Syndrome

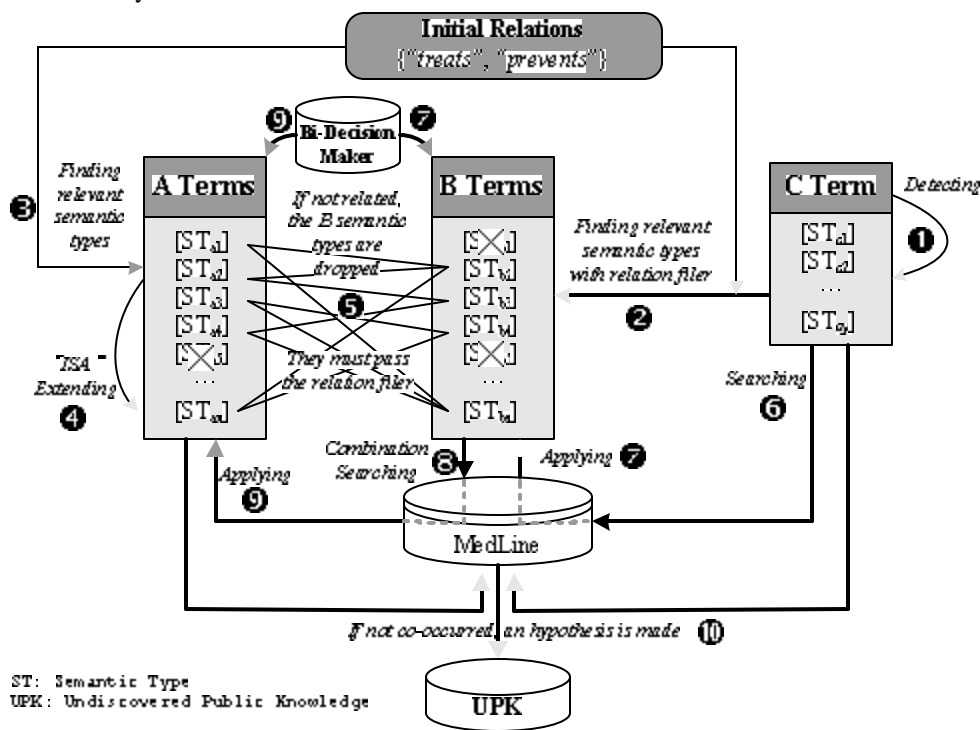


Figure 3. The data flow of Bio-SbKDS

Step 3: In order to derive the semantic types of A terms, the initial semantic relations (e.g. *treats*, *prevents*) are used. Here, it is important that the C term is set as “a subject” or “an object” for the initial relations. If the term is set as an object, only the semantic types on the first (not third) column in the Table 2 are considered in the search space.

Table 2. Semantic relations for some semantic types

Semantic Types (as subjects)	Relation	Semantic Types (as objects)
Antibiotic	<i>treats</i>	Disease or Syndrome
Drug Delivery Device	<i>treats</i>	Disease or Syndrome
Medical Device (too general)	<i>treats</i>	Disease or Syndrome
Pharmacologic Substance	<i>treats</i>	Disease or Syndrome
Therapeutic or Preventive Procedure	<i>treats</i>	Disease or Syndrome

However, if a semantic type is too general, then that type is ignored. Whether or not a semantic type is “too general” is decided by its hierarchy level. Currently Level 1, 2, 3 (e.g. A1.4.1) in the UMLS semantic network are regarded as “too general” because the terms in the semantic types in such levels are too broad.

Step 4: Extend the semantic types identified in STEP 3 by following through the ISA relations. Also “too general” semantic types are ignored. Actually through this process all sub-semantic types of the semantic types in STEP 3 are added to the semantic type list. For example, because [*Antibiotic*] is a sub-semantic type of [*Pharmacologic Substance*], [*Antibiotic*] is added. The four semantic types in STEP 3 are extended to 15 types through this process. These semantic types (*ST_A_can_ext*) are used for the semantic types of A terms as a category restriction.

Step 5: Because there must exist at least one relationship between A terms and B terms, Bio-SbKDS should check if there is at least one relationship between *ST_B* (the semantic types for B concepts in STEP 2) and *ST_A_can_ext* (the semantic types for A concepts obtained in STEP 4). First, for each semantic type for B terms Bio-SbKDS checks if there exists at least one relationship with any of the semantic types of A terms. If a semantic type for B terms does not have any relationship with any of the semantic types of A terms, the semantic type is dropped from the semantic type list of B terms. After this process is done with the semantic types of B terms, the same process is performed for the semantic types of A terms. These processes are called *mutual qualification*. During the mutual qualification procedure, Bio-SbKDS simultaneously checks if the two semantic type sets (for A terms and B terms) pass the predefined relation filter between A terms and B terms. These filter relations are *interacts_with*, *produces*, and *complicates*. Table 3 shows the two semantic type sets for B concepts and A concepts that are automatically generated using only the initial relations and the relation filters. However, other methods [2][6][8][9] manually generate them and use the same semantic type sets for B and A concepts.

Table 3. The Semantic Types as Category Restrictions for B Concepts and A Concepts

A Concepts	B Concepts
Indicator, Reagent, or Diagnostic Aid	Cell Function
Antibiotic	Carbohydrate
Biologically Active Substance	Eicosanoid
Pharmacologic Substance	Steroid
Chemical Viewed	Mental or Behavioral Dysfunction
	Element, Ion, or Isotope

Functionally Immunologic Factor Receptor	Organophosphorus Compound
Biomedical or Dental Material	Congenital Abnormality
Therapeutic or Preventive Procedure	Amino Acid, Peptide, or Protein
Vitamin	Organism Function
Hormone	Pathologic Function
Enzyme	Organ or Tissue Function
Hazardous or Poisonous Substance	Chemical Viewed Structurally
Neuroreactive Substance or Biogenic Amine	Nucleic Acid, Nucleoside, or Nucleotide
	Organic Chemical
	Cell or Molecular Dysfunction
	Inorganic Chemical
	Acquired Abnormality
	Molecular Function
	Neoplastic Process
	Mental Process
	Genetic Function
	Lipid
	Experimental Model of Disease
	Physiologic Function

Step 6: In order to collect B term candidates, the starting concept C is searched against Medline. Here, we should consider what B terms should be. Because there should be some meaningful semantic relationships between B terms and C term (for B terms to be a bridge between A terms and C term), B terms should be the major topics (concepts) of the documents by the keyword searching against Medline. Therefore, we collect only MajorTopic MeSH terms from the downloaded documents and calculate their counts. The rationale to consider the counts of B candidates here is that we try to find something (as A concepts) that is strongly associated with C concepts.

Step 7: B term category restrictions, which consist of semantic types obtained in STEP 5, are applied to the MeSH terms extracted in STEP 6. Also “too general” MeSH terms are excluded. The top N terms are selected as B concepts (Currently, N is 5).

Step 8: Unlike the initial search based on the starting concept C in STEP 6, Bio-SbKDS searches all top B terms against Medline. The B terms are ranked by the counts of the terms. On searching, the same date range is used as the initial keyword. However, the documents, relevant to C concept should be excluded. Thus, the search keyword would be “B term AND Date_Range NOT C term”. Similar to STEP 6 only MajorTopic

MeSH terms are collected. A sample keyword to be searched is the following:

"Blood Viscosity"[MAJOR] 1983[dp]:1985[dp] NOT "Raynaud Disease"[MeSH]

Step 9: A term category restrictions, which consist of the semantic types obtained in STEP 5, are applied to the MeSH terms extracted in STEP 8. Also "too general" MeSH terms are excluded. In addition to those qualifications, Bi-Decision Maker [5] determines if the MeSH terms are appropriate to A concepts. Through these processes, A concept candidates are generated.

Step 10: Because we try to find only novel C-A relationships, the system eliminates A candidates that already have some relationships with C concept by searching Medline; if C and A concepts co-occur together in the biomedical literature, those A concepts are dropped from the candidate list. From the A candidates, the top N_a as A concepts are selected based on their weights from the B term.

4. Experimental Results

In our experiments, we reimplemented two existing approaches: LSI-based [2, 6, 7,13] and association-rule based [1,8] for mining the hidden links and compared them with our Bio-SbKDS on two of Swanson's famous medical discoveries, "Raynaud Disease – Fish Oils" and "Migraine – Magnesium".

4.1 LSI-based Algorithm The particular "latent semantic indexing" (LSI) analysis that we have tried uses 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 [7]. For any matrix X (m by n), it can be decomposed as three matrices $T_0 S_0 D_0^T$. The method easily project T ($\geq k$) dimension and D ($\geq k$) project to the same k space. K is minimum rank of matrix T and D. Position in the space then serves as the new kind of semantic indexing. In this case, we can keep term vectors and document vectors point the directions about where they point, while we successfully reduce the dimension the data. If we compare two terms, we compare the two row vectors from the matrix $T_0 \times S_0$ since $XX^T = T_0 S_0^2 T_0^T$. The cosine between two row vectors reflects the extent to which two terms have a similar pattern. By cosine measure, the larger the value is, the more similar the two terms are. Theoretically, this will achieve a better comparison than that of standard cosine measure on original matrix. 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" (C) as input term, we can calculate the closeness of all the other terms after SVD analysis. Then we choose those top ranked terms (B) as input terms, thereby we can rank all the other terms (A) that are disjointed with "Raynaud disease".

The procedure of LSI algorithm:

- (1) Download k documents from PubMed through concept "C term" query
- (2) Extract all the terms as B terms from Meshheadinglist, 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
- (5) For each B_i ($i=1, 2, 3, \dots, 100$) 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

Association Rule-based Algorithm Association rules identify collections of data attributes that are statistically related in the underlying data. An association rule is of the form $B \rightarrow A$ where B and A are disjoint conjunctions of attribute-value pairs. Here we take C concept as input, and then we calculate all $B \rightarrow C$ rules. Then we generate all $A \rightarrow B$ rules. Last, we apply the transitive law to get the hidden link: $A \rightarrow C$. It must be noted that associate rule algorithm can not be applied to get $A \rightarrow C$ directly because A and C are not supposed to occur in the same data set. A and C are connected through the bridge concept B. For association rule, Support (B) indicates the probability that B occurs. Accordingly, Support (BnC) indicates the possibility that B and C occur together. Conf means confidence. $\text{Conf} = \text{Sup}(B \text{ n } C) / \text{sup}(B)$ indicates the confidence that B implies C ($B \rightarrow C$). $F(B \rightarrow C)$ measure is a measure of confidence of $B \rightarrow C$. The larger the value is, the more we are confident that $B \rightarrow C$.

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 B terms from MeSHheadinglist, 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
4. Generate all $B \rightarrow C$ association rules (B terms) and rank all the B terms according to $F = \frac{2 \text{Sup} \times \text{Conf}}{\text{Sup} + \text{Conf}}$ and then choose top n B terms.
5. For each B_i ($i=1, 2, 3, \dots, n$) do
 - a) Download k documents from PUBMED through “B Not C term” query within the same time period
 - b) Repeat step (2) to extract all candidate A terms
 - c) Repeat step (3), (4) to build term doc matrix after removing unrelated terms
 - d) Repeat step (5) to generate $A \rightarrow B$ rules (A terms)
6. List all $A \rightarrow B$ rules (A terms).

4.2 “Raynaud Disease – Fish Oils”

In the first experiment (table 4), the starting concept C is “Raynaud Disease”. Because we try to find something to “treat” or “prevent” the disease, we selected “treats” and “prevents” as the initial semantic relation. Using these initial semantic relations, the semantic types as category restrictions for B and A terms are generated. Our automatically-generated semantic types include most of the semantic types that [8] and [9] manually generated. While [8] and [9] used the same semantic types for both A and B terms, our model uses the different semantic types for B and A concepts because the roles of B and A concepts for C concept are different.

Table 4. Experiment results of “Raynaud Disease—Fish oils” problem (# of B=3 vs. # of B=5)

Top 3 B Concepts	Top 5 B Concepts
Blood Viscosity Quinazolines Pyridines	Blood Viscosity Quinazolines Pyridines Vinyl Chloride Imidazoles
Top 1 A Concept	Top 5 A Concepts
Fish Oils	“Anti-Inflammatory Agents, Non-Steroidal” Nicotine Niceritrol Antilipemic Agents

	<i>Fish Oils</i>
--	------------------

In our LSI-based experiments, the initial query is “Raynaud Disease [major] 1980:1985[edat]”. We ranked the top 100 B terms close to “Raynaud Disease”, from which we then submit each “B term 1980:1985[edat]” query. 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 MeshHeadingList or extracted both from MeshHeadingList and Title, Abstract, also according to the cell of matrix, term frequency (TF), term frequency and inverse document frequency (TFIDF), and Z-Score.

Table 5 LSI(Raynaud Disease—Fish Oil)

Selected Top B terms from which fish oil is discovered for each experiment	B term is the No. # Closest term to Raynaud disease (C)	Fish oil (A) is No. # closest term to B term	Term-Document Matrix representation
Plethysmography	17	766	MeSH TF
Arteriosclerosis	37	9253	
Eczema	41	1456	
Blood viscosity	70	300	MeSH TFIDF
Plethysmography	17	483	
Blood viscosity	70	2765	MeSH ZScore
Plethysmography	17	475	
Blood viscosity	70	442	MTAB TF
Arteriosclerosis	37	1693	
Eczema	41	1557	
Plethysmography	79	1466	MTAB TFIDF
Eczema	53	2440	
Arteriosclerosis	67	2097	MTAB ZScore
Eczema	52	1568	
Arteriosclerosis	67	1188	

In table 5, we only show some 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”, while A concept “fish oil” is the 475 closest to it. We can also see that measures TF, TFIDF and ZScore don’t affect results too much, while adding title and abstract to MeSH terms will 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, while 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 rank of each fish oil, we can see that LSI might not be a good method to discover A concept. Although it might get a little bit better result if we include some most frequently used MeSH terms in the stop list, it would not change the whole image of the ranking.

Table 6. Association rule (Raynaud Disease—Fish Oil)

Selected Top B terms from which fish oil is discovered for each experiment	Minimum # of B→C rules	Minimum # of A→B rules	Term-Document Matrix representation
Plethysmography	702	8731	MeSH TF
Plethysmography	702	7673	MeSH TFIDF
Plethysmography	2224	61330	MTAB TF
Platelet aggregation	2224	121818	MTAB TFIDF

In table 6, we made similar association rule experiments as LSI (please refer to the explanation of last table). Here we use $F = \frac{2 Sup \times Conf}{Sup + Conf}$ to measure the closeness between two terms such as A and B or B and C. From the experiment result in the table above, we can see that B term “plethysmography” is recognized as the first B term to generate A term fish oil by three experiments, the minimum

A→B and B→C rules of which are getting larger when adding title and abstract to mesh terms. However, TF and IDF do not affect result too much. Besides, we also found some other intermediate B terms. For example, in experiment MESH+TF, we found B terms such as prostaglandins E (44), blood viscosity (49), and platelet aggregation (54). These terms are all ranked within top 60.

4.3 “Migraine – Magnesium”

Swanson inferred the relationship between migraine and magnesium in 1988 [10]. He noticed that spreading cortical depression (B concept) is implicated in migraine (C concept) and also perceived that magnesium (A concept) inhibit spreading cortical depression (B concept).

Using Bio-SBKDS, it identifies that Cerebrovascular Circulation is strongly related to migraine (table 7). Actually Cerebrovascular Circulation is related to the spreading cortical depression. Therefore, we believe our results are very promising because our system finds out and ranks all the correct A concept in 4th and 15th. The terms in bold and italic, are those Swanson found manually.

Table 7. Experimental results of “Migraine—Magnesium” (# of B=3 vs. # of B=5)

Top 3 B Concepts	Top 5 B Concepts
“Cerebrovascular Circulation” Thiophenes <i>Ergotamines</i>	“Cerebrovascular Circulation” Thiophenes <i>Ergotamines</i> Platelet Aggregation <i>Propanolamines</i>
Top 4 A Concepts	Top 15 A Concepts
Protirelin Dihydroergotoxine Ergoloid Mesylates <i>Magnesium</i>	Protirelin Dihydroergotoxine Ergoloid Mesylates Tranlycypromine Dimethyl Sulfoxide Vinca Alkaloids Hydralazine Ephedrine “Cardiac Surgical Procedures” Orosomucoid “Extracorporeal Circulation”

Table 8 LSI (Migraine – Magnesium)

Selected Top B terms from which magnesium is discovered for each experiment	B term is the No. # Closest term to Migraine	Fish oil (A) is No. # closest term to B term	Term-Document Matrix representation

	(C)		
Ergolines	9	259	MeSH TF
Nicergoline	10	256	
Benzamides	13	866	
Pre-eclampsia	14	332	
Ergolines	9	282	MeSH TFIDF
Nicergoline	10	256	
Benzamides	13	822	
Blood Pre-eclampsia	14	424	
Ergolines	9	739	MeSH ZScore
Nicergoline	10	256	
Benzamides	13	689	
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

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—Magnesium”: 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 9 Association rule (Migraine – Magnesium)

Selected Top B terms from which magnesium is discovered for each experiment	Minimum # of B→C rules	Minimum # of A→B rules	Term-Document Matrix representation
Cerebrovascular circulation	671	5408	MeSH TF
Cerebrovascular circulation	671	5416	MeSH TFIDF
ergotamine	1923	14975	MTAB TF
Stress	1176	10402	MTAB TFIDF

Besides B concept in the table above, in the experiment MESH+TF, we also found B concepts such as muscle contraction, serotonin, food hypersensitivity, ischemic attack, calcium channel blockers, brain ischemia, aspirin, and spreading cortical depression (rank 49). All these terms are ranked within top 50.

Both experiments indicate that Bio-SbKDS generates fewer novel but relevant connections than the association rule based algorithm.

5. Conclusion and Future Work

This paper proposed a semantic based biomedical literature 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. 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/>].

6. Reference

- [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] 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*, 2003, 10(3):252-259.
- [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] Hu X., *Mining Novel Connections from Large Online Digital Library Using Biomedical Ontologie*, Library Management Journal, special issue in Libraries in the Knowledge Era: Exploiting the knowledge wealth for Semantic Web Technology, May 2005
- [6] Lindsay, R.K, and Gordon, M.D. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, 1999, 50(7):574-587.
- [7] 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
- [8] Pratt, Wanda and Yetisgen-Yildiz, Meliha, LitLinker: capturing connections across the biomedical literature, *K-CAP'03*, Sanibel Island, FL, Oct. 23-25, 2003 pp. 105-112.
- [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*, 1988, 31(4):526-557.
- [11] Swanson, DR. Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7-18, 1986.
- [12] Swanson, DR. Undiscovered public knowledge, *Libr. Q.*, 1986, 56(2):103-118.
- [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