MINING NOVEL KNOWLEDGE FROM BIOMEDICAL LITERATURE USING STATISTICAL MEASURES AND DOMAIN KNOWLEDGE

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Kishlay Jha

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

June 2016

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

MINING NOVEL KNOWLEDGE FROM BIOMEDICAL LITERATURE USING STATISTICAL MEASURES AND DOMAIN KNOWLEDGE

By

Kishlay Jha

The supervisory committee certifies that this thesis complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Wei Jin
Chair

Dr. Simone Ludwig

Dr. Na Gong

Approved:

06/16/2016
Date

Dr. Brian Slator
Department Chair

ABSTRACT

The problem of inferring novel knowledge from implicit facts by logically connecting independent fragments of literature is known as Literature Based Discovery(LBD). In LBD, to discover hidden links, it is important to determine the relevancy between concepts using appropriate information measures. In this study, to discover interesting and inherent links latent in large corpora, nine distinct methods, comprising variants of statistical information measures and derived semantic knowledge from domain ontology, are designed and compared. A series of experiments are performed and analyzed for those proposed methods. Also, a new strategy of effective preprocessing is proposed, which is capable of removing terms that have meager chances of constituting a new discovery. Finally, an organized list of final concepts deemed worthy of scientific investigation are provided to the user. Overall, our research presents a comprehensive analysis and perspective of how different statistical information measures and semantic knowledge affect the knowledge discovery procedure.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Wei Jin for trusting my abilities as a novice and providing me an opportunity to learn and explore as a part of her research team. I will always remain indebted to her persistent guidance and suggestions. I also take this opportunity to thank my committee members Dr. Simone Ludwig and Dr. Na gong for their insightful comments, feedbacks and moral encouragement.

Beside my committee members, i would like to thank my department, Computer Science, for providing me the best environment to work and study. Lastly, i remain grateful to everyone who directly or indirectly helped me accomplish this milestone.

DEDICATION

This thesis is dedicated to all

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

1. INTRODUCTION

Decades of experimentation and analysis has led to proliferation of scientific literature in the domain of biomedicine. MEDLINE, a preeminent bibliographic database contains more than 23 million references to journal articles in life science with a major concentration in biomedicine. Approximately 2,000-4,000 references are added everyday[17]. Figure 1.1 and 1.2 illustrates the exponential growth of scientific reporting and *Unified Medical Language*(biomedical knowledge base) provided Metathesaurus concepts over the past decade respectively. This overloaded textual resource in life science, although readily available, has made it difficult even for domain experts to subsume relevant knowledge in their field of interest. Sophisticated technologies and efficient linguistic computational tools are needed to leverage this rich representation to gain deeper insights. With the growth of this unparalleled publicly available scientific knowledge and availability of higher throughput methods, there has been a surge of interest in biomedical researchers to apply automated text analysis techniques and accelerate the discovery of new knowledge. This methodology of generating hitherto unknown but meaningful knowledge is known as Literature Based Discovery(LBD). In other words, LBD refers to the technique of harnessing already available scientific knowledge to uncover "non-apparent but interesting relationships" by rationally connecting complementary but non-interactive set of articles. LBD is considered a challenging aspect of biomedical text mining as it involves not only identification and extraction of information from text, but also logically connecting them to reveal hidden, complex and meaningful associations. Initially, Swanson and Smalheiser pioneered this area of research by studying the role of dietary fish oils in preventing Raynaud's syndrome (i.e., a vasospastic disorder causing the narrowing of blood vessels[27]). In subsequent years, to demonstrate the applicability of their ideas, they reported 11 previous unknown linking connections between migraine and magnesium[28]. Broadly, in their research, they found that implicit pieces of information could be discovered by studying the linkage between unrelated literature already present in the corpora. The principal objective behind this research arena was to identify plausible relations worthy of further scientific investigation or experimentation. Swanson classified LBD into two types, namely, open and closed discovery. In open discovery, a researcher specifies a topic of interest(viz., a disease or gene or pharmacological substance) and the system applies text mining techniques to find a set of terms that are directly related to the starting topic of interest(A). These terms are called intermediate terms(B) or *bridge concepts*. For each of these intermediate terms, the system reiterates the same mechanism to generate a set of terms that are directly correlated to each intermediate terms. Thus generated terms are called terminal concepts or final terms(C). It should be noted that the kind of connection between starting topic of interest and final terms are both indirect and novel. Generally, Open discovery relates to hypothesis generation, where none existed before. It begins with one concept from the research question and explores next levels(B and C) to identify relevant concepts which are unknown yet seemingly elucidate interesting associations. On the contrary, in closed discovery, the user specifies a pair of topics(A and C) and the objective is to find any unknown but meaningful connection that exists between them. It is more often characterized as hypothesis testing or generation of more granular hypotheses. A high level view of both open and closed discovery can be seen in figure 1.3. In this work, we restrict our discussion to open discovery.

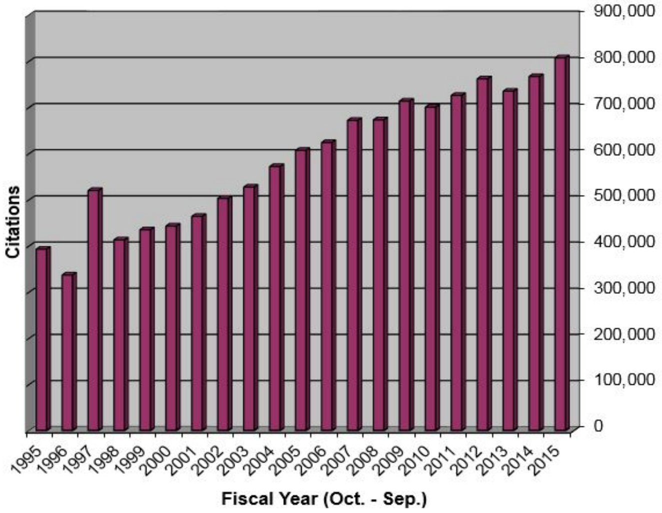


Figure 1.1. Number of indexed citations added to MEDLINE

Meanwhile, the initial works of Swanson and Smalheiser became prototypical example of LBD, it also simulated researchers to contribute in practical areas of protein-protein interaction, clinical medicine and health care. A few successful examples of LBD include, finding functional connection between genes[4], drug-disease association[30], identification of viruses as bioweapons[29]

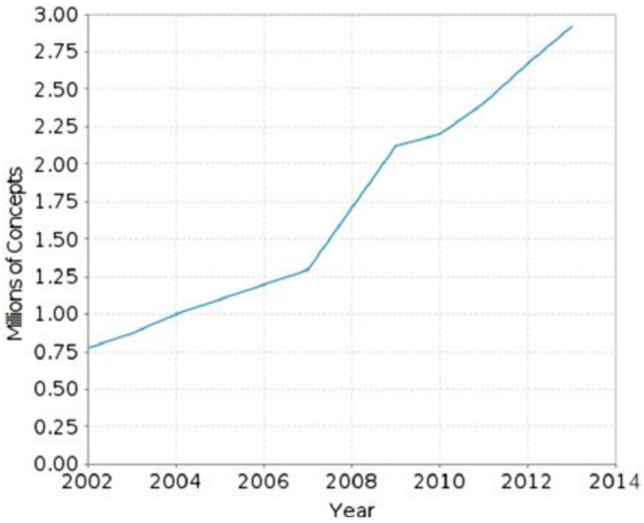


Figure 1.2. Growth of UML Metathesaurus each year

and so forth. Over the years, to tackle this intriguing problem, several techniques utilizing frequency of co-occurrences[9, 19, 26], association rules[14, 21] and graph-theoretic metrics[33, 3, 2] were proposed. Although these aforementioned methods immensely aided in developing scalable solutions and advanced the research area of LBD, there are possible areas of improvement. Prominent information scientists working in this area of study stress the need to improve upon the following issues a) investigating measures capable of generating related concepts(intermediate and terminal) with higher confidence, i.e., terms which are not only statistically prominent but also semantically associated b) development of prudent ways to navigate the large search space and prune uninformative, bogus terms in advance c) lessen the amount of manual intervention or domain knowledge required during the discovery process. In this study, we intend to probe these problems by exploring the idea that *interesting links or connection which help to elucidate implicit associations are better explored by integrating statistical correlation measures and semantic knowledge in an intelligent way.* Obviously, to find interesting connections, an information measure is required to determine the closeness between two terms. In this work, we study nine methods of mining hidden links from biomedical literature which are combinations of information measures and semantic support. These nine methods are further classified into three groups. The first group consists of three existing information measures: association rule, mutual information, and Chi-Square. The second group includes null-invariant correlation measures: all_confidence, Kulcynski, and cosine. Finally, the last group is a combination of correlation measures and our proposed concept of semantic relatedness. To the best of our knowledge, we are the first to study the application of these popular null-invariant correlation measure in biomedical literature mining. Also, in addition to a comparative study of information measures, we perform an extensive preprocessing to remove terms which are highly frequent, common, and uninformative. Our experiments demonstrate as to how it aids in reducing the generation of uninteresting rules, ultimately, improving the overall performance. Finally, to reduce the need for any manual intervention or domain knowledge, we incorporated available semantic knowledge as an integral component of our system. Unlike other approaches[26, 21], we require our users to input only possible semantic relations between initial topic of interest(A) and to be discovered target concept(C), rather than manually providing probable semantic types for intermediate and target terms. With input semantic relations and initial topic of interest, we automatically generate semantic types and use them as category restriction for B and C terms.

Overall, our research work presents a meticulous analysis of how manifold statistical information measures and semantic knowledge affect the knowledge discovery procedure. The reminder of this thesis is structured as follows. Chapter 2 discusses related work. In chapter 3, we present an overview of our methods in detail. In chapter 4, we present experiments and evaluation results. And finally chapter 5 brings conclusion and gives directions for future work.

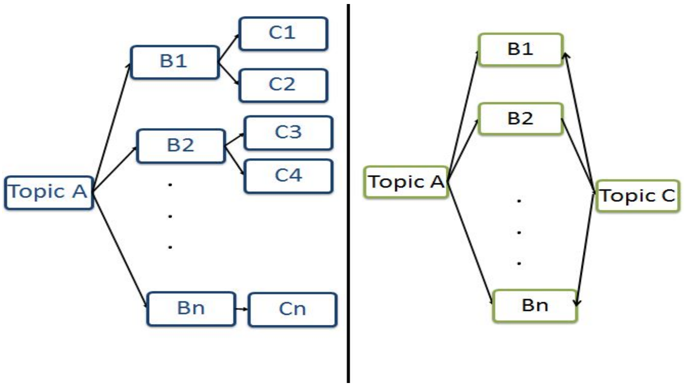


Figure 1.3. Open and Closed discovery approach

2. RELATED WORK

The original conception of LBD was facilitated by "Raynaud's disease-Fish oil" discovery by Swanson in 1986[27]. They proposed a simple ABC model (See Figure 2.1), where AB and BC refer to the direct relationships reported in literature explicitly wherein the goal was to find any inferred relationship via intermediates B. This model was used to propose several novel hypotheses by manually connecting missing links between disjoint journal articles in biomedical domain. For instance, in his famous "Raynaud's disease-Fish oil" discovery, he studied the literature related to Raynaud's disease and observed that patients with Raynaud's syndrome have high platelet aggregation, high blood viscosity and impaired vascular reactivity. On the other hand, the literature related to Fish oils contained information that ingestion of fish oils lowered blood viscosity, platelet aggregation, and caused vascular reactivity. Thus, by connecting these disjoint sets of literature, he hypothesized that fish oils may be beneficial for patients with Raynaud's syndrome. Likewise, in another study, using the same approach he found 11 indirect connections between Magnesium and Migraine Disorder, some of which are: serotonin, epilepsy, spreading cortical depression, calcium channel blockers, prostagladins, inflammation, type A personality and brain hypoxia. These connections was later verified experimentally by [22]. In subsequent efforts, together with Smalheiser, he postulated several other discoveries including Estrogen-Alzheimers Disease [24], Indomethacin-Alzheimers Disease [23] and Calcium Independent Phospholipase A2-Schizophrenia [25].

2.1. Arrowsmith

Although Swanson's initial investigation was based on exhaustively reading title and abstracts from MEDLINE, in years followed, he developed a software tool named *Arrowsmith* to automate some of the steps. It supported both open and closed discovery. Given an initial topic of interest(A and C), firstly, it queried the titles of MEDLINE articles belonging to both set of citations in order to generate an initial set of intermediates. Next, using the initial set of intermediates, MEDLINE was queried again to obtain more documents from which potentially useful B terms were obtained. These intermediates found were then ranked on the basis of frequency of co-occurrence. Later on, additional features like semantic filtering by Unified Medical Language[1]

[1] https://www.nlm.nih.gov/pubs/factsheets/umls.html

semantic types and more sophisticated ranking of intermediate terms were incorporated. Even though his work instituted seminal ideas in this area of study, it had a few setbacks. One of the major setbacks was the need for manual inspection of literature and domain knowledge required during several stages of discovery process. Consequently, subsequent works tried to alleviate this bottleneck by automating the process.

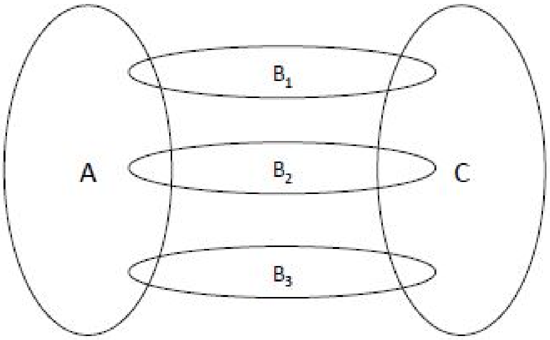


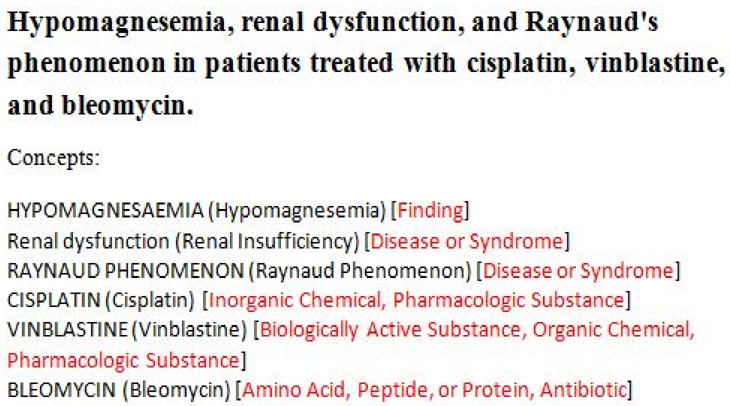
Figure 2.1. Swanson's ABC Model

2.2. Dad

Weeber et al[31] developed a concept based LBD system named Dad using Metamap[2]. Metamap is a tool developed by National Library of Medicine(NLM) to provide access to concepts in the UMLS Metathesarus from biomedical text. This is a powerful tool used by bioinformatics community to leverage the available domain knowledge. An example of concept extraction by Metamap is shown in Figure 2.2. In the figure, terms within big brackets marked as red are the corresponding semantic types. Weeber used these semantic types for concept filtering. The use semantic types also provides a better understanding of context from the hierarchical and associative relations in the semantic networks. Using concept based approach with aid of available domain knowledge, he successfully replicated some of Swanson's discoveries and also found some potentially new applications for thalidomide[32]. They suggested that thalidomide, through some immunologic factors such as tumor necrosis factor and interleukin-12, might be useful for treating acute pancreatitis, chronic hepatitis C, Helicobacter pylori-induced gastritis, and myasthenia gravis. The aforementioned approaches were based on the traditional understanding that discoveries are

[2] https://metamap.nlm.nih.gov/

likely to emerge from logical connection between initial topic of interest(A), intermediates(B) and terminal concepts(C) which frequently or rarely co-occur with each other in the knowledge base. Thus, building upon this idea, several distribution approaches[19, 26, 9] employed frequency based metrics such as term-inverse document frequency(*tf-idf*), record frequency and token frequency to find intermediate and terminal concepts.



Hypomagnesemia, renal dysfunction, and Raynaud's phenomenon in patients treated with cisplatin, vinblastine, and bleomycin.

Concepts:

HYPOMAGNESAEMIA (Hypomagnesemia) [Finding]
Renal dysfunction (Renal Insufficiency) [Disease or Syndrome]
RAYNAUD PHENOMENON (Raynaud Phenomenon) [Disease or Syndrome]
CISPLATIN (Cisplatin) [Inorganic Chemical, Pharmacologic Substance]
VINBLASTINE (Vinblastine) [Biologically Active Substance, Organic Chemical, Pharmacologic Substance]
BLEOMYCIN (Bleomycin) [Amino Acid, Peptide, or Protein, Antibiotic]

Figure 2.2. A sample biomedical title with concepts parsed by Metamap

2.3. Litlinker

While frequency based approaches were successful in propelling LBD one step ahead, there were certain issues remaining to be addressed. One of them was the possible number of $A \rightarrow B$, $B \rightarrow C$ combinations. Obviously, because in MEDLINE, one concept may be connected to many other concepts. Hence, it was necessary to explore solutions which can navigate such large search space in an efficient manner. To deal with this combinatorial problem, Pratt and Yetisgen-Yildiz[21] in their work 'Litlinker' used Unified Medical Language(UML) provided domain knowledge to limit their search space. They implemented open discovery building upon the initial framework established by Swanson. In addition to using knowledge base as an integral component, they grouped together synonym terms by merging any terms that had same concept id, and then assigned a preferred name to that group. Overall, the system included knowledge based methodologies, natural language processing techniques, and a data mining algorithm to mine biomedical literature for potentially casual links between biomedical terms. To identify correlated concepts, they used Associations rules(Apriori algorithm) and level of support to rank AB and BC term pairs. The threshold for support was empirically set to 0.002 which in their system meant, an association was likely to be spurious unless the concept occurred in at least five titles. In their later work[36], they reported the use of UML concepts is computationally expensive for practical use and decided to use MeSH[3] terms to represent documents. To reduce search space, they decided to prune non interesting intermediate or target terms. In their study to judge the significance of terms they report three class of problems on the basis of which they eliminate terms a) terms that were too broad (e.g. adults, disease, and medicine) to be target terms b) terms which were closely related to start term (e.g. headache for starting term migraine) c) terms that didn't make sense for plausible connections. For the first class of problem, they utilized MeSH hierarchy. In MeSH hierarchy, terms are ordered from generic to specific, any target term which were more generic than one or more linking terms were eliminated. Next, for the second class of problem, they again made use of MeSH hierarchy. Any term which were immediate family(e.g. parents, grandparents, siblings, children) of start terms were removed. Lastly, for the third class of problem they required user to select semantic types for intermediate linking and target terms. Any term which did not belong to the selected semantic types were considered non-interesting. To identify correlated concepts, they used a statistical approach based on background distribution and term probabilities. Using this approach, they were able to recover intermediates for Migraine disorder-Magnesium and also suggest new insights into associations between 1) Alzheimer's Disease and Endocannabinoids, 2) Migraine and AMPA receptors, and 3) Schizophrenia and Secretin. In our present work, to manage the exponential combinations of $A \rightarrow B$ and $B \rightarrow C$, we perform an extensive preprocessing to remove frequent terms that are too general to be meaningful.

2.4. Iridescent

Following the notion of distributional approaches, Wren et al[34] in IRIDESCENT, attempted to extend the calculation of mutual information to indirect associations by using Mutual Information measure of the shared associations. Given their maximum likelihood estimates, the strength of associations between AB and BC pairs were then computed and normalized, based on degree centrality between terms. This aided to remove non-informative terms that were frequently co-occurring and highly connected in the corpus. Applying their approach, the authors demonstrated the discovery of new knowledge on Chlorpromazine and Cardiac Hypertrophy.

[3] https://www.nlm.nih.gov/mesh/

2.5. Manjal

Padmini[26] in 2004 presented another LBD system named 'Manjal' based on concept profiles consisting of weighted MeSH terms. Her system supported both open and closed discovery. She viewed Swanson's discovery method as having two dimensions. First dimension referred to a set of interesting related concepts for a particular topic of interest. Second dimension explored the nature of relationships that existed between identified associations(AB or BC). In her work, her proposed algorithms focused on the first dimension and the second dimension was performed through manual analysis of literature. The idea was to build MeSH based profiles from MEDLINE for given topic of interest. Here a profile is refereed to as a set of MeSH terms that together represent a corresponding topic. For instance, consider a topic such as *Diabetes Mellitus*, the profile for this topic would include terms representing proteins, genes, drugs, treatments, other disease and symptoms associated with it. The fact that each MeSH terms belongs to one or more semantic types is exploited. To elaborate, topic profiles are built within context of semantic types. Thus, when required, the profiles may be focused or narrow down to specific semantic types. A normalized weighting scheme of TF × IDF (term frequency × inverse document frequency) is used to weight MeSH descriptors. Open and closed discovery algorithm employing MeSH based profiles were proposed to generate novel hypotheses. The methodology was used to replicated 5 out of 6 discoveries made by Swanson. Later on, it was also used to gain new insights into the novel therapeutic roles of turmeric.

2.6. Rajolink

Unlike other approaches which heavily relied upon the idea of frequent terms, Petric developed a LBD system called Rajolink based on rare terms. They implemented Swanson's ABC model in a different way. The main distinguishing feature was the combination of open and closed process together for knowledge discovery. It also differed from existing approaches in the way it identified candidate sets for to-be discovered concepts. In their approach, the choice of to-be-discoverd concept(A) was based on rare terms identified in the literature of initial topic under investigation(C). The motivation behind was - if a piece of information appears rarely in a set of articles then they they assume it has been explored relatively lesser by researchers. Thus, building upon this assumption, the authors argue that investigating these terms might prove as innovative pathways. It is known that rare connections might prove as golden luggage in large datasets. The key components in Rajolink included: Rare terms, Joint terms, and linking terms. The authors applied their method on autism literature and suggested relations between calcineurin and autism. Although no direct evidence of calcineurin role in the autism was found, the authors claim to have identified significant links between them by analyzing the articles from two domains. Another plausible hypothesis given by them was the relation between autism and NF-kappaB.

2.7. Bitola

Hristovski applied associate rule mining to find correlated MeSH terms using Swanson's open discovery approach and developed a system called Bitola[14, 13]. Bitola supported both open and closed discovery. The entire MEDLINE database was preprocessed and transformed into a local knowledge-base consisting of concepts and associations. This knowledge base was further used as a foundation for entire approach. As the search and analysis of results were performed on this locally stored knowledge base, the overall performance of system was fast. To find correlated concepts, they used Association rules, together with support and confidence. Bitola made extensive use of domain knowledge provided by UMLS. Consequently, it has several filtering options available when searching for related concepts: by semantic type, semantic group and by relationship strength. Furthermore, some methodological and technical developments were added later on to make it better for the genetic application. In addition to MeSH terms, gene symbols were extracted from the titles and abstracts from MEDLINE. For genes, chromosomal locations were loaded, as well as the chromosomal locations for numerous genetic disease. Also, it incorporated supplementary concept records - mostly drugs and chemicals. Lastly, the final target terms could be filtered by chromosomal location and expression location.

Although co-occurrence based methods advanced the research area of LBD, it was not foolproof. The use of co-occurrence has several drawbacks a) all co-occurrence in MEDLINE were not necessarily interesting. b) systems tend to produce a large number false positives (semantically unrelated associations) c) users have to review article manually to understand the nature of associations. Thus, to alleviate these shortcomings, researchers introduced the idea of relation based approach for LBD. A relation based approach utilizes semantic relations or predicates between concepts to capture the meaning of associations. Hristovski[12] used this approach based on predications extracted from two Natural language processing components, SEMREP and BiOMEDLEE.

To exploit semantic predications in LBD, they introduced a notion of discovery pattern. The patterns were classified into two forms on the basis of manner they generated candidates. The first form was named as Maybe_Treats1 which was satisfied when there was a change in body substance(B) associated with starting disease(C) and there was an opposite change in concept B associated with concept A. An example of this pattern is Swanson's Raynaud's disease-Fish oils case. Patients with Raynaud's syndrome(C) suffered from increased level of blood viscosity(B) and the ingestion of fish oils(A) reduced its level. Thus, fish oil may treat Raynaud's disease. Another form of discovery pattern introduced was Maybe_treats2, where in order to find potentially new treatment for a starting disease(A), another disease(A2) with similar characteristics was found and then a new treatment(C2) for disease(A) was proposed. An example for this pattern was demonstrated using example of Huntington disease. In patients with Huntington disease, the level of insulin is often decreased which is also the case for patients with Diabetes Mellitus. With the help of clinicians, the authors reported potentially new treatment for huntington disease - insulin. Also, the authors state an interesting fact to support their assertion - Huntington patients develop diabetes mellitus about seven times more often than matched healthy controlled individuals. Similar to the Maybe_Treats discovery pattern used by Hristovski, [1] introduced another pattern based on semantic predications named May_Disrupt. The pattern is of the form Substance X <inhibits> Substance Y, Substance Y <causes> Pathology Z, Substance X <may_disrupt> Pathology Z. It focuses on understanding relationships among drugs, genes, and disease. The objective was to use discovery pattern to understand the mechanism underlying drug therapies that are currently used but poorly understood. The methodology was used to investigate antipsychotic agents used in treatment of cancer. In their results they suggest five biomolecules: brain-derived neurotrophic factor, CYP2D6, glucocorticoid receptor, PRL, and TNF which may provide casual links between anti-psychotic agents and cancer.

2.8. Graph Based Approaches

On contrary to above approaches, [3] implemented relation based technique for closed discovery using graph based approach. The main idea was to generate a ranked set of subgraphs which captured multifaceted complex associations, given a pair of initial concepts. The subgraphs generated on distinct thematic dimensions enabled broader understanding of the nature of complex associations between concepts. To create subgraphs they relied upon three datasets. First was MEDLINE, a bibliographic database of more than 23 millions citations maintained by National

Library of Medicine. The Second was SemMedDB, a database of more than 65 million semantic predications extracted from MEDLINE by SEMREP. The third was Biomedical Knowledge Repository(BKR), a knowledge base consisting of statements from the UML Metathesarus together with semantic predications. The overall approach was divided into five steps: 1) Query specification 2) Candidate graph generation 3) Path context representation 4) Path Clustering and 5) Subgraph ranking. The query specification step required two initial concept of interest (A,C), path length(K) and a date(D). The initial concepts were manually augmented with other closely related concepts. Next, in candidate graph generation step, the system retrieved the set of MEDLINE documents relevant to input query and created a graph. Depth first search (DFS) algorithm was used to perform traversal and generate all paths of specified length(K). In third step (i.e. Path context representation), MeSH terms are used to define context of a path. The related paths were clustered into subgraphs. Dice similarity was used to compute the semantic similarity of MeSH descriptions representing paths. The paths above certain threshold were grouped together using Hierarchical agglomerative clustering (HAC) algorithm. Finally, in the last step, the generated subgraphs were ranked using intra-cluster similarity. This approach facilitated the re-discovery of 8 out of 9 existing discoveries. In addition to re-discovery, a statistical evaluation was done to measure the interestingness of a subgraph in MEDLINE. Interestingness was measured using rarity.

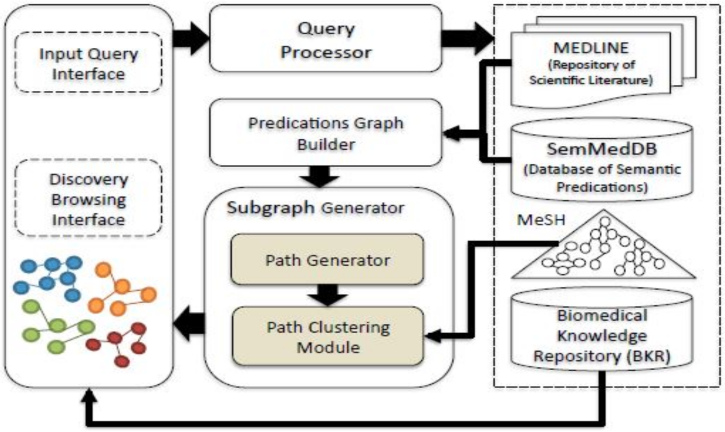

Input Query Interface
Query Processor
MEDLINE (Repository of Scientific Literature)
Predications Graph Builder
SemMedDB (Database of Semantic Predications)
Discovery Browsing Interface
Subgraph Generator
MeSH
Path Generator
Path Clustering Module
Biomedical Knowledge Repository (BKR)

Figure 2.3. System architecture of Cameron et al.

In [10] , Gramatica explored computational linguistics and graph theory to find new treatments for existing drugs. Leveraging upon publically available biomedical knowledge, they created a graph representation of knowledge to discover hidden relations between any drug and disease. The nodes in graph were referred to the UML concepts and links were defined by the co-occurrence of concepts at sentence-level. Analyzing the graph using stochastic process provided an effective instrument to understand different mechanisms of action of peptides and drugs. To shed light on the applicability of their technique, two examples were provided: a) the granulomatous disease Sarcoidosis and its pulmonary pathology, and b) Imatinib, a targeted-therapy agent against cancer cells, well known for its apoptosis action. Similarly, Goodwin et al[7] developed a hybrid approach using spreading activation, degree centrality, and relative frequencies for LBD. The approach generated a single subgraph by capturing the strength of associations between concepts. The strengths were calculated based on concept and predications based degree centrality. The spreading activation algorithm was then used to select relevant concepts. Finally, the system generated a list of intermediates instead of a graph. Overall, the method was used to successfully rediscover the connection in the Testosterone–Sleep discovery, and also elucidated the Norepinephrine, Depression, and Sleep scenario.

2.9. Bio-Sars

Again taking advantage of semantic relations, [15] proposed a biomedical semantic based association rule system(Bio-Sars) to generate highly likely novel and biomedically relevant connections among the biomedical concepts. The main distinctive characteristic was the augmentation of traditional association rule mining with semantic support to reduce associations which were spurious, useless or biomedically irrelevant. The relation based approach utilized semantic types, semantic relations and semantic hierarchy on the bridge and target concepts to filter out meaningless association rules. They also introduced a concept of mutual qualification. In mutual qualification, if semantic types of A and B for a rule ($A \rightarrow B$) are not related, the rule is dropped. In their work, the authors demonstrate the significance of mutual qualification in association rule to generate more semantically meaningful relations. The experiments replicated two of famous Swanson discoveries: Raynaud's disease-Fish Oils and Magnesium-Migraine Disorder. Moreover, they reimplemented Latent semantic indexing algorithm to compare their results. Although these were made significant strides they still required a certain amount of domain knowledge in order to specify appropriate semantic types for generating intermediate and terminal concepts. In this work, we automate this step by automatically generating semantic types for intermediate and terminal concepts by utilizing user provided “initial topic of interest(A)” and “initial semantic relations” for to-be-discovered concept. Closely related to our work is [15, 16, 18], where we generate the semantic types in a similar way but are distinct in a sense that we do not limit the use of semantic types to merely remove uninteresting relations. Instead, we go one step ahead by calculating the semantic relatedness between semantic types over MEDLINE corpus and use that information to promote relations with higher semantic meanings. Undoubtedly, these aforementioned works have furthered Swanson’s method significantly but none of them were comprehensive enough in evaluating various information measures and consider specific semantic relationships. Also, a limitation of measures such as Chi-Square and MI is that they suffer from a critical property of *null-invariance* (i.e., measures which are influenced by total number of null transactions). Null transaction in the context of biomedical dataset refers to articles not containing the concepts(A,B or B,C) of interest being examined. And as studied in[11], a good information measure should not be affected by transactions that do not contain the itemsets of interest, as it might generate unstable results. Motivated with this narrative, we were interested in studying the application of null-invariant correlation measures such as *all_confidence, Kulcynski and cosine* in the biomedical dataset and see how they affect the experiments results in comparison to existing methods. We believe, we are among the first to study the application of null-invariant measures in biomedical literature and present a comprehensive comparative study of how different information measures combined with semantic knowledge affect knowledge discovery process.

Although the basic foundation of LBD paradigm has been ABC model (initially proposed by Swanson), it still might miss some interesting A-C connections. Wilkowski et al. [33] extended this model using a graph based approach. Wilkowski suggested that the ABC model can be decomposed to a more granular level in order to elucidate complex associations between concepts. The extension proposed is known as AnC model, where,

$$\text{n}= (B_1, B_2, ..., B_m).$$

The main goal was to incorporate semantic predications and graph based techniques to elucidate understanding of poorly understood associations by providing novel viewpoints, observed upon expanding B element of ABC model. Similar to previous methods, semantic predications were extracted from MEDLINE citations using SEMREP. Also, it is worth nothing that while majority of studies in LBD focused on biomedical domain, there were few which explored other domains. Gordon et al. (2002)[8] applied Swanson's ABC model to discover novel applications for existing problem solutions on the World Wide Web. For instance, they used "genetic algorithms" as their A term and discovered many potential fields of application such as "virtual reality", "computer graphics", and "fluid dynamics". Cory (1997) applied LBD on humanities databases to discover hidden analogies.

3. METHODS

In this section, we present nine methods(categorized into three groups) to mine implicit associations from biomedical literature. The first group includes traditionally used information measures such as: Associate rule mining(ARM), Mutual Information, and Chi-square. The second group consists of popular null-invariant correlation measures such as: all_confidence, Kulcynski, and cosine. Finally, the third group includes combination of null-invariant measures and our proposed concept of semantic relatedness. A high level view of our methods is shown in figure 3.1. Also, a detailed algorithm for each method is presented.

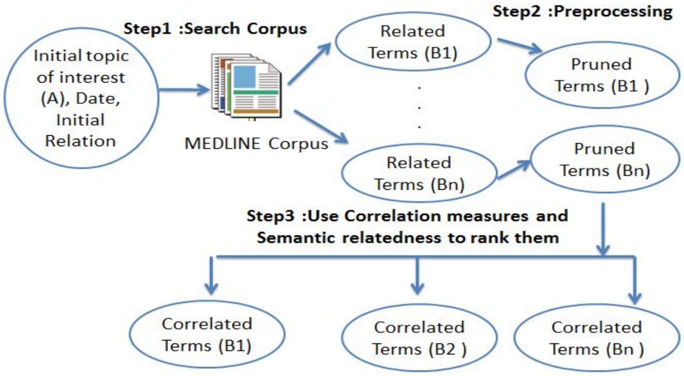


Figure 3.1. Basic architecture of the proposed methods

3.1. Group 1: Information Measures

3.1.1. Associate Rule Mining

Associate Rule Mining(ARM) is widely used in data mining applications. Given a document collection, specific level of support and confidence, the goal is to generate frequent itemsets above these thresholds. An association rule of the form $A \to B$, let $sup = support(A \cap B)$ and $conf = support(A \cap B)/support(A)$. If concept A is taken as input, then all $A \to B$ rules are found from one itemset. Then from another distinct itemset $B \to C$ rules are found. Finally, a transitive law is applied to get a transitive link $A \to C$. It should be noted that we can not find $A \to C$ directly, as both A and C occur in independent itemsets. In our experiments, we use F-measure(F) to calculate the strength of relation.

$$F = \frac{2Sup * \times Conf}{Sup + Conf}$$

3.1.2. Mutual Information

Mutual Information(MI) is used to measure the dependency between variables or terms. The degree of closeness is used to rank terms. For a given term pair(A,B), mutual information is computed as

$$MI(A, B) = log(P_{AB}/P_A.P_B)$$

where P_A,P_B denote the probability of term A and B respectively. P_{AB} denote the probability that terms A and B co-occur. In our experiments, to avoid negative weighting, we remove the log function. Also, it is worth nothing that this metric might rank rare associations higher[34].

3.1.3. Chi-Square

Given two variables, Chi-Square(χ^2) can measure how strongly one variable implies the other, based on the available data. For example: Suppose a pair (a,b), χ^2 takes into the account co-occurrence frequency of a,b and also co-occurrence of a and b with other terms. For a term co-occurrence matrix, let O be the observed frequency and E be the expected frequency, then the χ^2 value is computed as

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

For, a 2 × 2 contingency table(shown in table 3.1, the degree of freedom is (2-1)(2-1) = 1. For 1 degree of freedom, the χ^2 value needed to reject null hypothesis at the 1% significance level is 6.63. In other words, if the Chi-Square value between two terms is greater than the critical value of 6.63, that means it rejects the null hypothesis that two terms are independent.

3.2. Group 2: Null Invariant Correlation Measures

3.2.1. All_confidence

Given a pair of terms, A and B, the all_confidence(all_conf) measure of A and B is defined as:

$$all\_conf(A, B) = support(A \cup B)/\max\{support(A), support(B)\}$$

Where max{support(A),support(B)} is the maximum support of itemsets A and B.

Table 3.1. 2 × 2 contingency table

	V=v	$V \neq v$
U=u	E_{11}=($R_1 \times C_1$)/N	E_{12}=($R_1 \times C_2$)/N
$U \neq u$	E_{21}=($R_2 \times C_1$)/N	E_{22}=($R_2 \times C_2$)/N

3.2.2. Kulczynski

For a pair of terms, A and B, the Kulczynski(Kulc) measure of A and B is defined as:

$$Kulc(A, B) = 1/2(P(A|B) + P(B|A)).$$

Kulc is a measure of average of two conditional probabilities: the probability of itemset B given A, and the probability of itemset A given B.

3.2.3. Cosine

For a pair of terms, A and B, the Cosine measure of A and B is defined as:

$$Cosine(A, B) = \frac{support(A \cup B)}{\sqrt{sup(A) * sup(B)}}$$

The reason all of these above three measures are called null-invariant is that their values are only influenced by A,B and $(A \cap B)$ and not by total number of transactions not containing A or B.

In our experiments, similar to [36, 26], we also use MeSH terms to represent articles. MeSH terms are National Library of Medicine(NLM) controlled vocabulary which human experts use to manually index citations. Thus, it is assertive to assume that if an article is important to a particular MeSH term, it will be indexed with that. Basically, MeSH terms are classified into three types: main headings (also known as descriptors), sub-headings(qualifiers) and supplementary concept records. Descriptors indicate the main contents of the citation. For illustration, if an article discusses about the role of fish oil in treating patients with Raynauds disease, then the article may be indexed with descriptors "fish oil","raynaud disease","blood vessels". At this point of writing, there are 27,883 descriptors. Moreover, if a descriptor alone is the central topic of article, it is assigned an attribute called "major topic". Another classification of MeSH term is Qualifer. But, qualifiers are important only when in conjunction with descriptor(i.e., they describe a special aspect of descriptor). Lastly, Supplementary concept records are used to index chemicals, drugs, and other concepts related to citation. In this work, we restrict our analysis to descriptors. Next, we present algorithms using methods contained in groups 1 and 2.

Algorithm

Input: Initial topic of investigation A as MeSH term, Date, K (top B concepts), M (top C concepts), Semantic relation for B and Semantic relation for C.

Output: Final concept list (C terms)

Procedure

- *Step 1.* Search the local MEDLINE database[Detailed in section 3.3.1] to find documents indexed with the input query MeSH term before the specified cut-off date.
- *Step 2.* Extract all the MeSH descriptors which co-occur with input MeSH term from relevant documents. We call these terms *all_B_Terms*.
- *Step 3.* Remove all terms from *all_B_terms* which belong to *common_MeSH_terms* set created in section 3.3.2. Also, remove terms which do not belong to generated semantic types for B(section 3.3.3). The remaining terms are *pruned_all_B_Terms*.
- *Step 4.* From the local database, find the co-occurrence frequency between A and each of candidate B terms.
- *Step 5.* Use statistical information measure(e.g. χ^2, MI or Kulc) to determine the closeness between terms.
- *Step 6.* Rank all the candidate B terms based on the degree of closeness. Select top K B terms.
- *Step 7.* For each *B_i* (i=1,2,3...K) do
    1. Search local MEDLINE database to find documents which are indexed with B but not A with the same cut-off date as Step 1.
    2. Repeat from Step 2 to Step 6 to generate candidate C terms.
    3. Remove all C terms co-occurring with A term. Select top M C terms.

- *Step 8.* List all C terms (Final terms)

Next, we explain each step in detail with an example. The input parameters are: Initial topic of investigation A (Raynaud's disease), Date (1985), K (10), M (1), Semantic relation for B ('Causes') and Semantic relation for C ('Treats').

- *Step 1.* Given the initial topic of interest, which at this point of time should be a MeSH descriptor, the local MEDLINE database[section 3.3.1] is searched to find all documents indexed with that term. Only those documents are retrieved whose publication is before cut-off date (e.g. 1985). We find 2646 documents.

- *Step 2.* All the MeSH descriptors which co-occur with Rayanud's disease(RD) are extracted from the retrieved documents. These terms are called *all_B_Terms*. The total terms B terms found were 2533.

- *Step 3.* To prune general terms, we remove all terms from *all_B_terms* which belonged to *common_MeSH_terms* set created in section 3.3.2. After that we find the semantic types of RD from UML semantic network. The semantic type for RD is "Disease or syndrome". Next, again leveraging semantic network, we find all semantic types which have relation 'causes' with Disease or Syndrome. The semantic types found are used as category restriction for B terms. Also, we find all semantic types which have relation 'treats' with Disease or Syndrome. These are used as category restriction for A terms. Table 3.3 shows the semantic type for A and B terms. All terms from *all_B_terms* whose semantic type does not belong to the semantic types generated for B terms are removed. The remaining terms are *pruned_all_B_Terms*. The total terms left after preprocessing were 957.

- *Step 4.* From the local database, we find the frequency of A, B and AB. For an example of association Raynaud's disease → Blood viscosity, we find frequency of Raynaud's disease (A):2646, Blood viscosity(B): 3911, and "Raynaud's disease-Blood viscosity(AB)": 30.

- *Step 5.* The frequencies found in step 4 are plugged into statistical information measure(e.g. χ^2, MI or Kulc) to determine the closeness between terms.

- *Step 6.* The B terms are ranked in descending order of their frequency. Top 10 B terms are selected to find terms for next level.

- *Step 7.* Each B terms is search in local MEDLINE database to find relevant documents. Similar to search in step 1, we retrieve all documents published before specified cut-off data (1985) but differ in a way that documents are indexed with B and NOT C term. This constraint of search 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. The terms hence extracted will be candidate A terms.

- *Step 8.* The terms in candidate A term set which belong to *common_MeSH_terms* set created in step 3 are removed. Likewise, any term whose semantic type doesn't belong to semantic types of A term generated in step 3 are removed.

- *Step 9.* Similar to step 4, we find the co-occurrence frequency of B→A associations. For an example term pair Blood Viscosity → Fish Oils, the frequency of Blood Viscosity(B) is 3911, the frequency of fish oils(A) is 860 and frequency of "Blood Viscosity-Fish oils" is 7. Next, plug in these values into statistical measures and calculate the score. This score represents the degree of closeness between B and A.

- *Step 10.* The final A terms for each top 10 B terms are ranked in descending order of frequency and top 1 term is selected. Altogether, final top 10 A terms are shown to the user.

3.3. Group 3: Combination of Null-invariant Measures and Semantic Support

In this section, we present our new method to discover novel knowledge from biomedical literature. The fundamental idea is to augment the method described in section 3.2 with semantic support. Basically, in this method, a user is required to specify an topic of investigation(A), initial semantic relation(ISR) and a date. For instance, if a user is interested in finding novel therapeutic preventions for Raynaud's disease, then the input parameters could be following, initial topic of investigation(A) "Raynaud's disease", date "1985", and ISRs "causes" and "prevents".

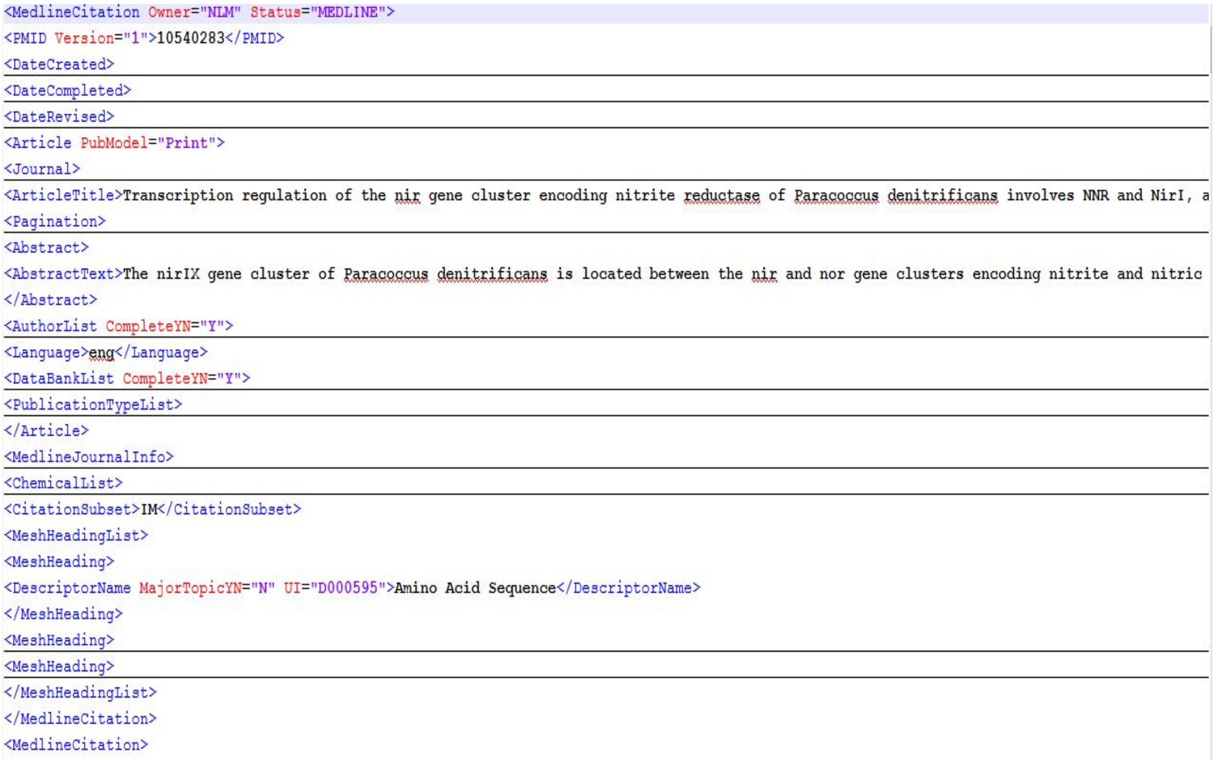
After the user specifies input parameters, our system performs a search on local database to collect relevant literature. Next, we perform an extensive preprocessing to eliminate terms which are highly frequent. To determine highly frequent terms, firstly, we calculate frequency of MeSH terms over entire MEDLINE corpus and draw a box plot[20] to find outliers. We assume that the outliers generated are highly common terms(refer section 3.3.2). Also, we take advantage of MeSH hierarchy to prune terms which are generic. Followed by preprocessing, our system automatically generates the semantic types using semantic network[4] for intermediate and final concepts. The semantic types are generated from user provided initial topic of interest and initial semantic relation. We used these generated semantic types as category restriction for intermediate(B) and terminal(C) concepts. A detail of explanation of this step is presented in section 3.3.3.

In addition to taking advantage of available semantic and category knowledge, we introduce a concept of semantic co-occurrence. To elaborate, similar to co-occurrence matrix at term level, we project MeSH terms to their semantic space and generate a co-occurrence matrix of semantic types (viz., the weighted matrix provides the count of a semantic type co-occurring with all other semantic types over the MEDLINE corpus). For instance, for a term pair (Raynaud's disease → Platelet Adhesiveness), we first obtain their respective semantic types (Disease or Syndrome → Cell function). Next, from the weighted semantic co-occurrence matrix, we find the frequency of co-occurrence between them. We use this value as a measure for our semantic relatedness. We assume that if semantic co-occurrence of two semantic types is high, then terms belonging to them are more related. It should be noted that each MeSH term certainly belongs to one or more semantic types. Altogether, combination of null-invariant measures and semantic co-occurrence value is used to measure the degree of closeness between terms(A→B, B→C). In essence, the idea is to promote relations which are both statistically significant and semantically associated. A detailed algorithm of this method is presented in section 3.3.3.

3.3.1. Searching the Literature

For searching the literature, we created our own local MEDLINE database. This database consists of entire dump of MEDLINE citation records(year 2015). The available raw data is in XML format. A sample medline citation in XML format is shown in figure 3.2. The raw data processed and stored across several tables. For each MEDLINE record, we store $PMID$(a single element to uniquely identify articles), $ArticleTitle$(the title of each article),$Abstract$(abstract text of each article),$PubDate$ (it contains the full date on which the article was published) and $MeshHeadingList$(it contains the MeSH terms assigned for each article). In our experiments, we use $PubDate$ to divide MEDLINE into two sets(before publication date and after publication date) for evaluation purposes. We also store MeSH tree codes for each MeSH term. The database

[4] https://semanticnetwork.nlm.nih.gov/



```
<MedlineCitation Owner="NLM" Status="MEDLINE">
<PMID Version="1">10540283</PMID>
<DateCreated>
<DateCompleted>
<DateRevised>
<Article PubModel="Print">
<Journal>
<ArticleTitle>Transcription regulation of the nir gene cluster encoding nitrite reductase of Paracoccus denitrificans involves NNR and NirI, a
<Pagination>
<Abstract>
<AbstractText>The nirIX gene cluster of Paracoccus denitrificans is located between the nir and nor gene clusters encoding nitrite and nitric
</Abstract>
<AuthorList CompleteYN="Y">
<Language>eng</Language>
<DataBankList CompleteYN="Y">
<PublicationTypeList>
</Article>
<MedlineJournalInfo>
<ChemicalList>
<CitationSubset>IM</CitationSubset>
<MeshHeadingList>
<MeshHeading>
<DescriptorName MajorTopicYN="N" UI="D000595">Amino Acid Sequence</DescriptorName>
</MeshHeading>
<MeshHeading>
<MeshHeading>
</MeshHeadingList>
</MedlineCitation>
<MedlineCitation>
```

Figure 3.2. Xml structure of sample MEDLINE file

design takes into account the peculiarities of MeSH Terms, the fact that there can be more than one MeSH tree code for one MeSH term. For instance, a MeSH term *Eye* has MeSH tree codes A01.456.505.420, A09.371 respectively. Also, it should be noted that we use this database to find the co-occurrence frequency between two terms. This value is used in calculation of several statistical information measures.

To summarize, in our system, a researcher is required to specify an initial topic of interest which should be a MeSH descriptor(e.g. "Migraine Disorders") and a cut-off date(e.g. 1988) to collect relevant literature on a particular subject of interest.

3.3.2. Preprocessing

As discussed in section 2.1, one of the major challenges for LBD researchers has been to enhance performance of their system by negotiating the exponential search space in an intuitive way. The general convention has been to remove terms which are highly "common". [21] initially removed terms or concepts which appeared more than 10,000 times in MEDLINE documents. Later, they used MeSH hierarchy(Tree codes) to remove terms which were too "broad". Similarly, [15] created their own custom stop word list to remove terms which they deemed unsuitable for discovery. This

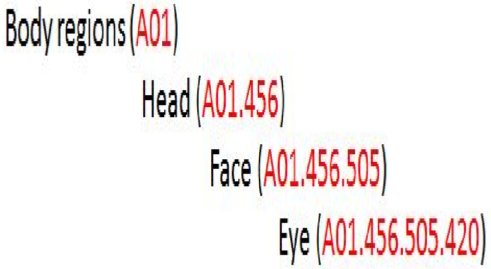


Figure 3.3. An example showing MeSH term hierarchy

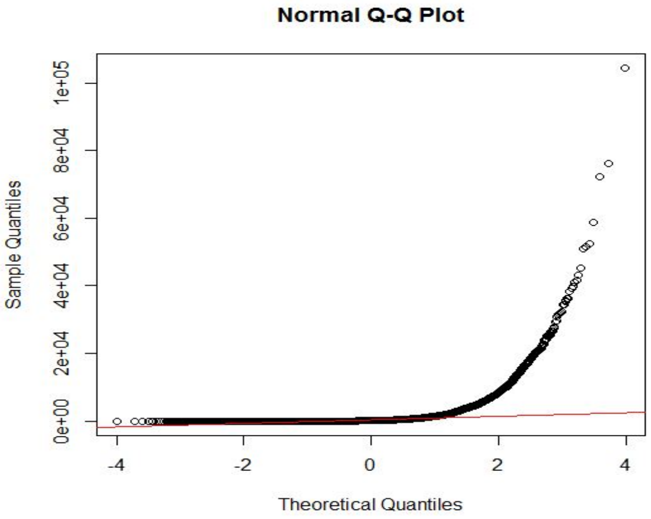


Figure 3.4. A normal Quantile-Quantile Plot

list included 325 frequently used MeSH terms. However, it is not clear, what parameters they use to deem a MeSH term as common. In our work, after studying the existing techniques and taking into account the statistical and semantic properties of MeSH terms, we decided to prune common terms based on following two parameters: a) frequency of MeSH terms over entire MEDLINE records b) tree codes of MeSH terms.

Firstly, after obtaining the frequency of MeSH terms over the entire MEDLINE corpus, we plot its distribution. To understand the nature of distribution, we drew a *Normal Q-Q plot*[5]. Generally, in the Normal Q-Q plot, for normally distributed data, the data points approximately fit a straight line. However, as it is evident from figure 3.4, the distribution in our case is not normal. Alternatively, it is highly skewed. And in statistics, for a dataset which does not follow gaussian or

[5] http://data.library.virginia.edu/understanding-q-q-plots/

Table 3.2. Top 10 common MeSH terms

MeSH term	Frequency	Major Count
Humans	12944044	1
Male	6326498	1
Female	6306642	0
Animals	5130327	10
Adult	3792522	441
Middle Aged	3171765	633
Time Factors	966947	1085
Child, Preschool	725299	547
United States	711204	0
Molecular Sequence Data	593668	39

normal distribution, Median is a preferred measure for central tendency[20]. Thus, to find outlier data, we draw a boxplot and obtain its outer fences.

$$UpperOuterFence : Q_U + 3(InterQuartileRange)$$

Measurements which lie beyond these outer fences are considered as outliers[20]. The upper outer fence value calculated was 24,404. Thus, any MeSH term with frequency greater than this value was considered as highly frequent a.k.a "common". In addition to outlier detection, we also take advantage of MeSH term hierarchy. MeSH terms are arranged in hierarchy according to their level of specificity, the term in the top are generic whereas terms in lower levels are more specific. To eliminate generic terms, we remove MeSH terms whose level is 1,2,3(e.g. A01, A01.456, A01.456.313). An example of MeSH hierarchy is shown in Figure 3.3. In total, using this technique we gathered 454 terms. We name this set as *common_MeSH_terms*. Table 3.2 shows top 10 common terms. It is interesting to note that the outlier terms obtained have very low support as major topics. For instance, terms like "humans","male" were assigned as major topic only once. This is encouraging because a term is assigned as "major topic" only when it is the central focus of article. And the outlier terms in our set having low support as major topic signifies their hollowness to produce a novel discovery. To summarize, as our goal in this research is to evaluate methods on the basis of their novelty in generating knowledge, we prune terms which have meager statistical or semantic significance.

3.3.3. Generating Semantic Types for Intermediate and Terminal Concepts

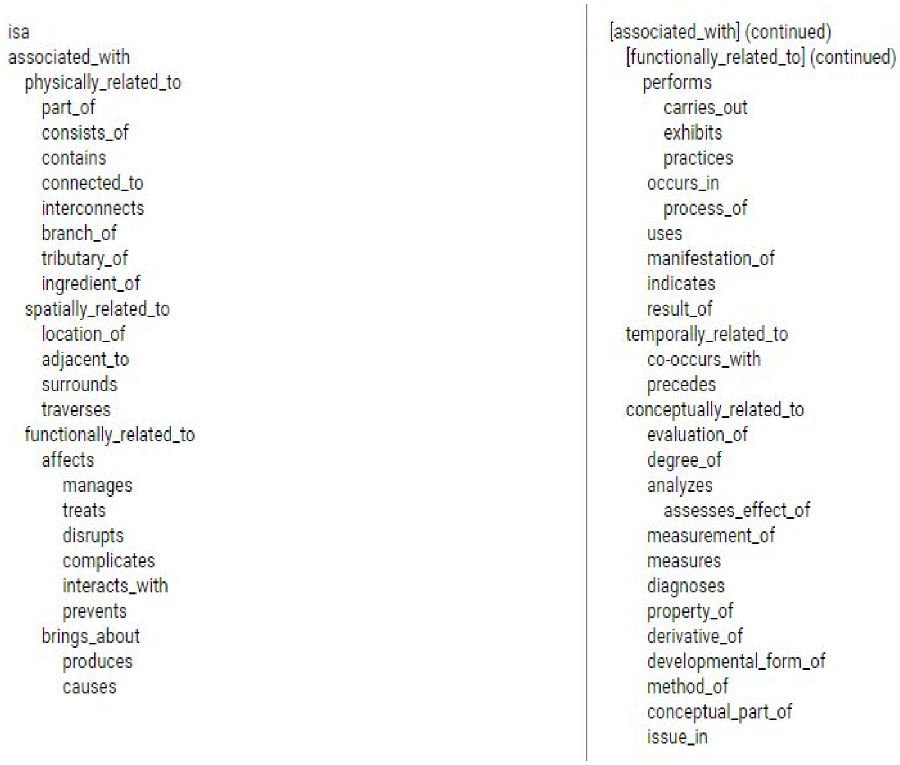
Given an initial topic of interest(A) and initial semantic relations(ISR), we use available domain ontology provided by UMLS to find semantic types for intermediate and terminal concepts. UML is a biomedical knowledge base and is used as an integral component throughout our knowledge discovery procedure. Primarily, it has three components a) **Metathesaurus**: It is a multi-purpose vocabulary database that is organized by concept, or meaning. It links alternative names and views of the same concept from different source vocabularies and identifies useful relationships between different concepts. b) **Semantic network**: All concepts in the UML metathesaurus are categorized into one or broader subject categories called semantic types. Ex:- Fish oil belongs to semantic types [“biologically active substance”, “lipid”, “pharmacologic substance”] and Raynaud’s disease belongs to semantic type [“Disease or syndrome”]. There are altogether 135 semantic types and there exists a set of useful relationships between them which are called “semantic relations”. At present, there are 54 semantic relations (See figure 3.5) between semantic types. Examples of relations includes “treats”, “diagnoses”, “prevents”, and so forth. In our methods, the user is required to input one of these semantic relations as an input parameter. c) **Specialist lexicon**:- The SPECIALIST Lexicon provides the word usage information needed for the SPECIALIST Natural Language Processing (NLP) System. The Lexicon entry for each word or term contains the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System.

For an input topic of interest(A) as “Raynaud’ disease” and ISRs (“causes”, “treats”), where ”causes” AND ”treats” refer to the constraints set for intermediate B and final C terms respectively, we first find semantic types for input A. For Raynaud’s disease, the semantic type is “Disease or Syndrome”. Next, we use the semantic network to find all the semantic types which have relations ‘causes’ or ‘treats’ with “Disease or syndrome”. Table 3.3 shows the semantic types generated for intermediate and final concepts.

Unlike other approaches[26, 32, 19], we automatically generate the semantic types for B and C terms instead of requiring users to manually set them. It should be noted that the semantic types manually set by [26] are automatically generated by our system. Also, we use a distinct set of semantic types for B and C terms rather than using the same for both. The semantic types for B and C are shown in table 3.3. Next, we present the algorithm for methods using group 3 measures.

Table 3.3. Semantic types for intermediate and final concepts for “Raynaud’s disease”

Semantic type for B	Semantic type for C
Physiologic Function	Therapeutic or Preventive Procedure
Organism Function	Chemical Viewed Functionally
Pathologic Function	Neuroreactive Substance or Biogenic Amine
Molecular Function	Biologically Active Substance
Organ or Tissue Function	Pharmacologic Substance
Genetic Function	Antibiotic
Pharmacologic Substance	Indicator, Reagent, or Diagnostic Aid
Biologically Active Substance	Immunologic Factor
Chemical Viewed Structurally	Hazardous or Poisonous Substance
Neoplastic Process	Indicator, Reagent, or Diagnostic Aid
Cell Function	Enzyme
Disease or Syndrome	
Cell or Molecular Dysfunction	
Element, Ion, or Isotope	
Amino Acid, Peptide, or Protein	
Antibiotic	
Cell or Molecular Dysfunction	
Nucleic Acid, Nucleoside, or Nucleotide	
Congenital Abnormality	
Acquired Abnormality	
Mental or Behavioral Dysfunction	
Mental Process	



```
isa
associated_with
  physically_related_to
    part_of
    consists_of
    contains
    connected_to
    interconnects
    branch_of
    tributary_of
    ingredient_of
  spatially_related_to
    location_of
    adjacent_to
    surrounds
    traverses
  functionally_related_to
    affects
      manages
      treats
      disrupts
      complicates
      interacts_with
      prevents
    brings_about
      produces
      causes
```

```
[associated_with] (continued)
  [functionally_related_to] (continued)
    performs
      carries_out
      exhibits
      practices
    occurs_in
      process_of
    uses
    manifestation_of
    indicates
    result_of
  temporally_related_to
    co-occurs_with
    precedes
  conceptually_related_to
    evaluation_of
    degree_of
    analyzes
      assesses_effect_of
    measurement_of
    measures
    diagnoses
    property_of
    derivative_of
    developmental_form_of
    method_of
    conceptual_part_of
    issue_in
```

Figure 3.5. Semantic relations available from semantic network

Algorithm

Input: Initial topic of investigation A as MeSH term, Date, K (top B concepts), M (top C concepts), Semantic relation for B and Semantic relation for C.

Output: Final concept list (C terms)

Procedure

- *Step 1.* Search local MEDLINE database to find documents indexed with the input query MeSH term before specified cut-off date.

- *Step 2.* Extract all the MeSH descriptors which co-occur with input MeSH term from relevant documents. We call these terms *all_B_Terms*.

- *Step 3.* Remove terms from *all_B_terms* which are present in *common_MeSH_terms* set created in section 3.3.2. The remaining terms are *pruned_all_B_Terms*.

- *Step 4.* Find the semantic type of term A from UMLS (Sem_A) and generate semantic types for intermediate term as explained section 3.3.3. These are referred to as Sem_B and are used as category restriction for B terms.

- *Step 5.* Remove all terms from *pruned_all_B_Terms* whose semantic types do not belong to Sem_B. We call these terms candidate B terms.

- *Step 6.* Use null-invariant measures to calculate the statistical value(*stat_value*) between A and each of candidate B terms. To obtain semantic co-occurrence value, we use semantic co-occurrence matrix(*sem_coccur*). Find cumulative value for each B term (i.e., Score (A→B1) = stat_value(A→B1) * sem_coccur (A→B1).

- *Step 7.* Rank all B terms in descending order of the overall score. Select top k B terms.

- *Step 8.* Similar to Step 6, find all related semantic types for C terms. These are called Sem_C and are used as category restriction for C terms.

- *Step 9.* For each *B_i* (i=1,2,3...k) do

    1. Search local MEDLINE database to find documents which are indexed with B but not A with the same date as Step 1.
    2. Repeat step 2 to step 6 to find C terms.
    3. Remove all C terms co-occurring with A term. Rank them in descending order of the overall score. Select top M C terms.

- *Step 10.* List all C terms (Final terms)

Next, we explain each step in detail with an example. The input parameters are: Initial topic of investigation A (Migraine disorder), Date (1988), K (10), M (1), Semantic relation for B (Causes) and Semantic relation for C (Treats).

- *Step 1.* Given the initial topic of interest, which at this point of time is restricted to MeSH descriptor, the local MEDLINE database[section 3.3.1] is searched to find all documents indexed with that term. Only those documents are retrieved whose publication is before cut-off date (e.g. 1988). We find 6116 documents.

- *Step 2.* All the MeSH descriptors which co-occur with Migraine disorder(MD) are extracted from the retrieved documents. These terms are called *all_B_Terms*. The total B terms found are 3643.

- *Step 3.* To prune general terms, we remove all terms from *all_B_terms* which belonged to *common_MeSH_terms* set created in section 3.3.2. After that we find the semantic types of MD from UML semantic network. The semantic type for MD is "Disease or syndrome". Next, again leveraging semantic network, we find all semantic types which have relation 'causes' with Disease or Syndrome. The semantic types found are used as category restriction for B terms. Also, we find all semantic types which have relation 'treats' with Disease or Syndrome. These are used as category restriction for A terms. Table 3.3 shows the semantic type for A and B terms. All terms from *all_B_terms* whose semantic type doesn't belong to the semantic types generated for B terms are removed. The remaining terms are *pruned_all_B_Terms*. The total terms left after preprocessing were 3424.

- *Step 4.* From the local database, we find the frequency of A, B and AB. For an example of association Migraine disorder → epilepsy, we find frequency of Migraine disorder (A):2646, epilepsy(B): 46884, and "Migraine disorder-epilepsy(AB)": 374.

- *Step 5.* The frequencies found in step 4 are plugged into statistical information measure(e.g. χ^2, MI or Kulc) to determine the closeness between terms.

- *Step 6.* The B terms are ranked in descending order of their frequency. Top 10 B terms are selected to find terms for next level.

- *Step 7.* Each B terms is search in local MEDLINE database to find relevant documents. Similar to search in step 1, we retrieve all documents published before specified cut-off data (1988) but differ in a way that documents are indexed with B and NOT C term. This constraint of search 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. The terms hence extracted will be candidate A terms.

- *Step 8.* The terms in candidate A term set which belong to *common_MeSH_terms* set created in step 3 are removed. Likewise, any term whose semantic type doesn't belong to semantic types of A term generated in step 3 are removed.

- *Step 9.* Similar to step 4, we find the co-occurrence frequency of B→A associations. For an example term pair epilepsy → Magnesium, the frequency of epilepsy(B) is 3911, the frequency of Magnesium(A) is 36914 and frequency of “epilepsy-Magnesium” is 212. Next, plug in these values into statistical measures and calculate the score. This score represents the degree of closeness between B and A.

- *Step 10.* The final A terms for each top 10 B terms are ranked in descending order of frequency and top 1 term is selected. Altogether, final top 10 A terms are shown to the user.

In our experiments, we empirically set the value of K as 10 and M as 1.

4. EXPERIMENTS

Evaluating LBD systems is an essentially challenging issue and remains an open problem[37]. Although LBD systems are designed to produce novel scientific knowledge, replicating Swanson's discovery has been seen as an effective evaluation approach by most LBD researchers. Swanson and Smalheiser applied their famous ABC model and published several discoveries in medical domain. Since then, their discoveries have become gold standard for evaluation. To compare and contrast our manifold methods, we choose two of Swanson's famous discoveries

1. Raynaud's Disease - Fish oil (RD-FO)
2. Migraine disorder - Magnesium (MD-MG)

In our experiments, we intend to explore the following questions:

1. How does the use of existing information measures such as "Associate rule mining, Mutual information, Chi-Square" compare with popular null-invariant measures in their application to biomedical dataset?
2. How does our proposed approach of augmenting null invariant correlation measures with semantic support affect experimental results ?
3. Does the preprocessing performed to remove general, uninformative links aid to improve the overall performance ?
4. Finally, Are the final C terms generated by different methods worthy of further scientific research or experimentation ?

4.1. Result of Raynaud's Disease - Fish Oils example

In 1986, Swanson explored the research question of "role of dietary fish oils in treating patients with Raynaud's syndrome". After analyzing disjoint sets of literature belonging to Fish oils and Raynaud's disease respectively, he found that Raynaud's disease is aggravated by high *blood viscosity*(B), high *platelet aggregation*(B), *Vasoconstriction*(B), and the ingestion of Fish oils

Table 4.1. Top 5 B terms for RD-FO before preprocessing

existing information measures			null-invariant correlation measures		
ARM	MI	χ^2	all_conf	kulc	cosine
humans	humans	female	scleroderma, systemic	humans	scleroderma, systemic
female	female	male	age factors	female	fingers
male	male	scleroderma, systemic	sympa thectomy	male	age factors
adult	adult	animals	fingers	adult	sympa thectomy
middle aged	middle aged	fingers	telan giectasis	middle aged	vibration

Table 4.2. Ranking of top 5 B terms for RD-FO after preprocessing

Existing information measures			Null-invariant Correlation measures		
ARM	MI	χ^2	all_conf	kulc	cosine
thromboangiitis obliterans	blood pressure	thromboangiitis obliterans	thromboangiitis obliterans	cervical rib	thromboangiitis obliterans
regional blood flow	pregnancy	arteriosclerosis obliterans	arteriosclerosis obliterans	acro-osteolysis	arteriosclerosis obliterans
arteriosclerosis	arthritis, rheumatoid	cryoglobulins	cryoglobulins	thromboangiitis obliterans	cryoglobulins
arteriosclerosis obliterans	arteriosclerosis	erythromelalgia	fingers	erythromelalgia	erythromelalgia
arthritis, rheumatoid	chronic disease	arteritis	intermittent claudication	chilblains	arteritis

Table 4.3. Ranking of important B terms in Raynaud's disease - Fish Oils before preprocessing

B term	Group 1			Group 2		
	m1	m2	m3	m4	m5	m6
Blood viscosity	61/2533	191/2533	51/1100	31/2533	80/2533	46/2533
Vasoconstriction	73/2533	262/2533	62/1100	30/2533	93/2533	53/2533
epoprostenol	140/2533	394/2533	190/1100	73/2533	176/2533	119/2533
thrombosis	90/2533	390/2533	154/1100	67/2533	190/2533	91/957
Platelet aggregation	238/2533	272/2533	500/1100	383/2533	413/2533	356/2533
arteriosclerosis	321/2533	154/2533	140/1100	134/2533	31/2533	95/2533

Table 4.4. Ranking of important B terms in Raynaud's disease-Fish Oils after preprocessing

B term	Group 1			Group 2			Group 3		
	m1	m2	m3	m4	m5	m6	m7	m8	m9
Blood viscosity	14/957	42/957	16/626	12/957	30/957	15/957	10/957	25/957	12/957
Vasoconstriction	18/957	62/957	25/626	16/957	36/957	11/957	11/957	35/957	10/957
epoprostenol	41/957	107/957	28/626	18/957	39/957	50/957	13/957	35/957	19/957
thrombosis	27/957	44/957	71/626	21/957	10/957	48/957	17/957	39/957	43/957
Platelet aggregation	73/957	65/957	140/626	349/957	173/957	148/957	193/957	214/957	208/957
arteriosclerosis	23/957	40/957	26/626	11/957	28/957	41/957	3/957	17/957	17/957

reduced these phenomena. Thus, he hypothesized that Fish oil(C) may be beneficial to people with Raynad's disease(A). Later on, it was clinically verified by Digiacomo in 1989[5].

To evaluate the performance of several methods, we conduct a series of experiments on MEDLINE data for this test case. In accordance with methodology, the experiments are also divided into three groups. The grouping of methods facilitates in providing a global picture on performance of diverse information measures on ranking implicit connections. Our readers should note that in tables 4.3, 4.4, 4.5 and 4.6, m1, m2, m3 in group 1 refer to Associate rule mining, Mutual information, and Chi-Square. Likewise, m4, m5, m6 in group 2 denote All_conf, Kulc, and Cosine. And m7, m8, m9 represent null-invariant measures supplemented with semantic support. Lastly, the fraction in cells of tables 4.3, 4.4, 4.5, and 4.6 is in the form of p_1/p_2, where p_1 denotes the rank of B terms and p_2 denotes the total number of A→B rules.

Before studying the comparison of different methods, we first discuss the role of preprocessing in eliminating common terms. To test our technique, we generated the intermediate concepts (B) for "Raynaud's disease" for first two groups(Group1, Group2) before and after preprocessing. Table 4.1 shows the top 5 ranked B terms without any preprocessing and Table 4.2 shows the top 5 after after preprocessing. It is evident to observe that the "common" MeSH terms ("humans", "male", "adult") which have meager chances of conceiving a novel discovery were ranked at high positions. It is encouraging to notice that these terms are present in our *common_MeSH_terms* set created in section 3.1.3.2. Also, it should be noted that the total number of rules generated without any preprocessing is 2533 (Table 4.3) ,whereas, after preprocessing it is reduced to 957 (Table 4.4). Overall, the ranks of important intermediate terms (Table 4.4) are boosted after preprocessing. The B terms in table 4.3 and 4.4 are the ones which lead to "fish oil" as final concept(C). Likewise, the connections mentioned by Swanson's paper are italicised. It is obvious to see that the frequent MeSH terms captured by our proposed *common_MeSH_terms* set dramatically reduce the rules generated and greatly boost significant B terms to higher ranks, which also demonstrates the importance of preprocessing step in improving knowledge discovery procedure.

Now, we discuss the results for different methods. For information measures in group 1, we found that mutual information preferred rare terms more but it could not rank important B terms better in comparison to the other two measures (Table 4.4). Interestingly, association rule provides relatively better ranking than Chi-square and Mutual Information. However, it generated more rules than Chi-Square. Obviously, as in Chi-square, if we remove any relationships with correlation value less than the critical value of 6.63, it would generate fewer rules. A careful observation elucidates that Chi-square was helpful in eliminating statistically insignificant terms. Next, for measures in group 2 (null-invariant measures), as illustrated in table 4.4, all three measures ranked most of the B terms better than information measures in group 1. We believe the better ranking for measures in this group is due to their null-invariant property(viz., they are not influenced by transactions which do not contain itemsets of interest). Also, recent studies tend to support this premise by suggesting that null-invariance is indeed a critical property for associations in large datasets[35]. Thus, a good information measure should not be influenced by null-transactions. Finally, for the third group, where we augment null-invariant measures with semantic relatedness, we notice that ranks for B terms are boosted. For instance, the ranks for *blood viscosity, vasoconstriction, epoprostenol* are improved. The improvement in ranks points out that the concept of semantic relatedness helps to promote terms which are more semantically meaningful.

4.2. Result of Migraine - Magnesium example

Swanson in 1988, proposed 11 previously unknown connections between Migraine disorder and Magnesium[28]. Some of them are epilepsy, serotonin, prostaglandins, substance p among others. It was later corroborated by Gallai[6]. Similar to FO-RD experiment, we examine our methods for this test case. Before we discuss results, we intend to aware our readers that for this particular test case, in MEDLINE, there were already a few articles before 1988 where Migraine disorder and Magnesium co-occurred(PMIDS : 3908832, 4922695, 7031826, 7031826). Therefore, in our experiment we exclude these articles from baseline dataset to prevent them from influencing our end results.

Table 4.5. Ranking of important B terms in the Migraine disorder - Magnesium query before preprocessing

B term	Group 1			Group 2		
	m1	m2	m3	m4	m5	m6
ergotamine	11/3645	147/3645	4/1533	3/3645	16/3645	4/3645
epilepsy	15/3645	30/3645	28/1533	13/3645	51/3645	24/3645
serotonin	17/3645	29/3645	31/1533	34/3645	49/3645	18/3645
caffeine	59/3645	137/3645	61/1533	19/3645	113/3645	57/3645
substance p	974/3645	1404/3645	-	746/3645	1277/3645	1004/3645
depression	57/3645	59/3645	292/1533	37/3645	56/3645	53/3645
nifedipine	895/3645	1266/3645	-	654/3645	1134/3645	870/3645

Table 4.6 shows the ranks of important intermediate terms (B) connecting Migraine and Magnesium. Much alike as in our previous case, we find significant improvement in ranks of B terms and reduction in the number of rules generated after preprocessing. Among information measures in group 1, as expected, mutual information again ranked rare terms better. However, in this scenario, we witnessed an important insight for Chi-Square. While Chi-Square undeniably generates lesser rules as compared to other information measures, it risks missing some important connections. For instance, in table 4.6, for terms *substance p*, *nifedipine*, Chi-square did not have any ranks because their scores were below the critical value(χ^2 less than 6.63). Thus, for terms which are important but have relatively low support in literature, χ^2 might risk missing them. On the other hand, null-invariant measures in group 2 again provides better ranking for most B terms including the ones missed by χ^2. The improved ranks by measures in group 2 manifest the significane for null-invariant property in information measures for large datasets. Lastly, for measures in group 3, we see reasonable improvement in ranks for important B terms(Table 4.6). Also, it is worthwhile to note that terms like *substance p, nifedipine* which have less support in literature were ranked better. Again, we believe the semantic support aided in boosting the ranks for terms which are more semantically related. Similar to previous test case, we calculate the overall score for each method.

4.3. Result Analysis and Discussion

To examine the precision of generated C terms for several methods, we divide the MEDLINE data into two sets: 1) a baseline set which includes citations before a selected cut-off date(i.e. input date from the user.) 2) a test set which includes publications after this specified cut-off date. We

Table 4.6. Ranking of important B terms in the Migraine disorder - Magnesium query after preprocessing

B term	Group 1			Group 2			Group 3		
	m1	m2	m3	m4	m5	m6	m7	m8	m9
ergotamine	6/3424	113/3424	4/1368	3/3424	12/3424	4/3424	2/3424	6/3424	2/3424
epilepsy	11/3424	19/3424	26/1368	9/3424	46/3424	21/3424	15/3424	22/3424	38/3424
serotonin	12/3424	18/3424	29/1368	11/3424	47/3424	25/3424	5/3424	89/3424	22/3424
caffeine	45/3424	105/3424	58/1368	19/3424	103/3424	50/3424	8/3424	58/3424	21/3424
substance p	866/3424	1223/3424	-	672/3424	1194/3424	1017/3424	785/3424	1029/3424	922/3424
depression	43/3424	39/3424	267/1368	36/3424	23/3424	47/3424	14/3424	23/3424	12/3424
nifedipine	792/3424	1097/3424	-	586/3424	1059/3424	799/3424	314/3424	583/3424	429/3424

implemented our methods on baseline set and checked the generated connections in the test set. To judge precision, we checked in our test set, whether the generated C terms appear with the start term(A) in the same citation. We assume that if A and C are mentioned in the same citation, they are related. In addition, as we restrict our C terms to Sem_C(i.e. semantic types for C terms which treat disease or syndrome), we assume that the terms in this set are possible treatments for a input disease. However, a drawback of this approach is that it cannot include relations that may appear in the future (viz., some of the target terms identified by our methods may become legitimate discoveries in the future but are considered incorrect target terms now as they do not appear together with the start term). In table 4.7 and 4,8, we show top 10 ranked C terms for methods in group 2 and group 3 respectively. In the brackets are the relevant PMIDs. To measure the precision of C terms, we consider terms which co-occur which with the start term in the same citation as true positives, and terms which are too general or do not co-occur with the start term as false positives. Overall, a unique score is calculated for terminal concepts belonging to each method. The score is calculated as sum of reciprocal ranks of relevant final terms(C) in the returned top 10 terms for each method. Mathematically, it is can be represented as below:

$$Score(m_t) = \sum_{i=1}^{10} \frac{1}{Rank(C_i)}$$

Where, m_t = {m1,m2,...,m9} and C_i refers to the relevant concepts. It is a reasonable measure of ranking method performance as it favours relevant terms that are ranked at higher positions while also giving appropriate weights to the lower ranked terms. Figure 4.1 illustrates the overall score of terminal concepts by methods. It should be noted that points in x-axis {1,2,...,9} denote methods {m1,m2,...,m9} respectively. The curves for two queries indicate that the overall score for methods in group 3 is greater than methods in group 2 which in turn greater than methods in group 1. We intend to highlight that some of the top ranked terms found for RD-FO test case in group 3 such as *'lipoproteins,vldl'*, *'niceritrol'*, *'platelet activating factor'* were also suggested by [26, 16, 15] in their top results. This provides an additional support to our proposed approach of augmenting null-invariant measures with semantic relatedness.

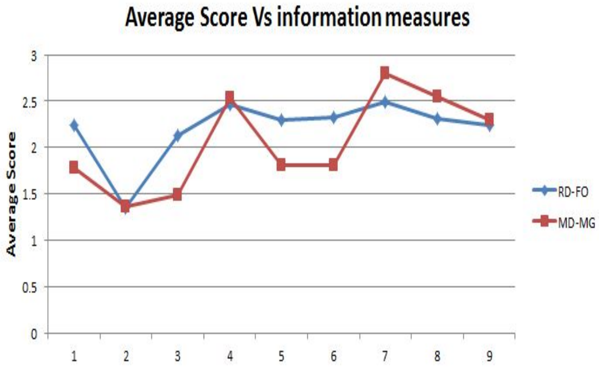


Figure 4.1. Average score for final concepts

Similarly, for MD-MG test case, figure 4.1 shows the overall score for all nine methods. The methods in group 3 had greater overall score in comparison to methods in group 1 and group 2. Also, in group 3, we find some important C terms, such as *'diet,sodium-restricted'*,*'phospholipases'*, *'amygdala'*, *'receptors, prostaglandin'* and so forth. Table 4.7, 4.8 shows the top 10 final concepts for methods in group 2 and group 3 respectively. We believe this catalogued list of final concepts(C) will help biomedical scientists to develop a cognitive perspective and analyze terms worthy of further scientific exploration.

Table 4.7. Final C terms for RD-FO and MD-MG in group 2

Raynaud's disease - Fish Oil			Migraine Disorder - Magnesium		
all_conf*SR	kulc*SR	cosine*SR	all_conf*SR	kulc*SR	cosine*SR
ligases (10959150)	tetrathionic acid (Not found)	microscopic angioscopy (25394956)	mandibular nerve (20618819)	chromans (11603382)	heparin, low-molecular-weight (19287274)
pityriasis (21807877)	pyroglobulins (16111177)	receptors, epoprostenol (1848945)	pargyline (8906292)	heparin, low-molecular-weight (19287274)	primidone (Not found)
pyroglobulins (16111177)	retinal artery (19171245)	hypohidrosis (10918257)	quipazine (Not found)	dibenzyl chlorethamine (Not found)	labor, induced (22280825)
ophthalmic artery (Not found)	receptors, epoprostenol (1848945)	pyroglobulins (16111177)	postpartum hemorrhage (12694520)	cinanserin (17351723)	pyrogallol (Not found)
receptors, thromboxane (1412196)	platelet activating factor[26]	hla-dr7 antigen (Not found)	mannitol (18953486)	phospho lipases (23826990)	cinanserin (17351723)
phospho diesterase inhibitors (25189168)	chlormezanone (Not found)	benz bromarone (Not found)	labor, induced (generic)	hypohidrosis (22492215)	tryptophan hydroxylase (24458851)
methacholine chloride (10959150)	niceritrol [16]	liver circulation (16724674)	desipramine (8712630)	patch tests (Not found)	ritanserin (9507121)
platelet factor 3 (Not found)	cystinosis (Not found)	amino acids, neutral (False)	tympanic membrane (12880669)	adenylyl cyclase inhibitors (1646776)	maxillary nerve (11797480)
pulmonary infarction (Not found)	ascorbic acid (12690904)	platelet activating factor[26]	amobarbital (10331688)	receptors, prostaglandin (24703233)	pergolide (19683643)
yersinia enterocolitica (Not found)	swine (Not found)	substantia gelatinosa (Not found)	mandibular nerve (20618819)	stereois omerism (26650258)	5,7-dihydroxy tryptamine (Not found)

Table 4.8. Final C terms for RD-FO and MD-MG in group 3

Raynaud's disease - Fish Oil			Migraine Disorder - Magnesium		
all_conf*SR	kulc*SR	cosine*SR	all_conf*SR	kulc*SR	cosine*SR
anesthesia, inhalation (7839003)	lymphokines (18571695)	cyclofenil (10796397)	diet,sodium-restricted (20713242)	tetrodotoxin (24292897)	doxepin (10436945)
lymphopenia (24294139)	receptors, epoprostenol (1848945)	phenazo pyridine (19300288)	dna replication (24266335)	betahistine (24166742)	maprotiline (14598505)
norethindrone (7875423)	receptors, epoprostenol (1848945)	lymphokines (18571695)	injections, intraventricular (25053746)	phospholipases (23826990)	isradipine (9812220)
dextro amphetamine (18431096)	fibrinogens, abnormal (12846071)	lipoprotein-x (Mentioned in [26])	amobarbital (10331688)	receptors, prostaglandin (24703233)	pergolide (19683643)
lipoproteins, vldl (Mentioned in[26])	niceritrol (Mentioned in [16])	pyroglobulins (16111177)	labor, induced (22280825)	audiometry, evoked response (15108495)	hypopituitarism (10524659)
tonometry, ocular (11879133)	ipoproteins, hdl3 (Mentioned in[26])	hrombasthenia (2244702)	cardiac output (25873813)	labetalol (12482217)	choroiditis (19220303)
phenyl thiazolylthiourea (Not found)	hyperalgesia (14569920)	rhinitis, vasomotor (Not found)	desipramine (8712630)	cinanserin (17351723)	quipazine (Not found)
amino acids, neutral (False)	antigens, human platelet (Not found)	lymphatic metastasis (2372019)	students, nursing (False)	methiothepin (10524657)	hypo pituitarism (10524659)
sradipine (9812220)	fibrinogens, abnormal (12846071)	cholesterol, dietary (False)	pargyline (8906292)	suicide (False)	isradipine (9812220)
lymphokines (18571695)	Antilipemic Agents (23347192)	sulindac (Not found)	njections, intraven tricular (25053746)	stomach neoplasms (20391683)	betahistine (24166742)

5. CONCLUSION

In this work, we compared nine different methods to generate novel knowledge from publicly available biomedical knowledge base. The methods were combinations of different statistical information measures and semantic support. Broadly, we classified them into three groups for better understanding of results. In addition to points raised in the study, we make the following particular contributions:

1. We performed a comparative study of several methods (combinations of statistical information measure and semantic support) and put forth a rationale behind how each of them affects results.
2. A notion of semantic relatedness was introduced and demonstrated as to how it assists in promoting semantically meaningful relations.
3. A new approach for extensive preprocessing was proposed to handle common MeSH terms. We perform statistical outlier detection and take advantage of MeSH hierarchy in this step. The experiments validate its utility.
4. We reduced the need for domain knowledge or manual interventation by automating the semantic types needed for intermediate and final concepts.
5. Finally, we generated an organized list of final C terms and provided references to PMIDs to assist medical researchers with further exploration.

To summarize our findings, an in-depth examination of diverse statistical information measures and semantic support reveals that different strategies favour certain types of concepts. In addition, as knowledge discovery is an open ended process, certain terms which are considered false positives at present may be realized as legitimate discoveries in the future. Thus, although evaluation of methods on Swanson's proposed discoveries brings into light some keen insights, it does not precisely illustrate what target terms should we emphasize most. Altogether, our experiments demonstrate that the best way to find meaningful final terms(C) is to rank them based on a combination of statistical information measures and semantic support drawn from domain ontology.

In future research, in addition to specific points raised in the study, we intend to add more semantic expressiveness to our generated hypotheses. We are looking at more specialized biomedical ontologies such as SEMREP[https://semrep.nlm.nih.gov/] for this purpose. Next, we intend to explore alternative measures such as Normalized goggle distance to calculate the degree of closeness between terms in our knowledge graph. Also, it would be interesting to see how random walk integrated with information measures affect the knowledge discovery process. Lastly, to strength our evaluation, we intend to investigate more robust evaluation techniques which not evaluate top ranked terms but the entire set of target terms.

REFERENCES

[1] Caroline B Ahlers, Dimitar Hristovski, Halil Kilicoglu, and Thomas C Rindflesch. Using the literature-based discovery paradigm to investigate drug mechanisms. In *AMIA*, 2007.

[2] David Cameron, Ramakanth Kavuluru, Olivier Bodenreider, Pablo N Mendes, and Amit P Sheth. Semantic predications for complex information needs in biomedical literature. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 512–519. IEEE, 2011.

[3] Delroy Cameron, Ramakanth Kavuluru, Thomas C Rindflesch, Amit P Sheth, Krishnaprasad Thirunarayan, and Olivier Bodenreider. Context-driven automatic subgraph creation for literature-based discovery. *Journal of biomedical informatics*, 54:141–157, 2015.

[4] Damien Chaussabel and Alan Sher. Mining microarray expression data by literature profiling. *Genome biology*, 3(10):1–16, 2002.

[5] Ralph A DiGiacomo, Joel M Kremer, and Dhiraj M Shah. Fish-oil dietary supplementation in patients with raynaud's phenomenon: a double-blind, controlled, prospective study. *The American journal of medicine*, 86(2):158–164, 1989.

[6] Virgilio Gallai, Paola Sarchielli, Giuliana Coata, Caterina Firenze, Piero Morucci, and Giuseppe Abbritti. Serum and salivary magnesium levels in migraine. results in a group of juvenile patients. *Headache: The Journal of Head and Face Pain*, 32(3):132–135, 1992.

[7] J Caleb Goodwin, Trevor Cohen, and Thomas Rindflesch. Discovery by scent: Discovery browsing system based on the information foraging theory. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pages 232–239. IEEE, 2012.

[8] Michael Gordon, Robert K Lindsay, and Weiguo Fan. Literature-based discovery on the world wide web. *ACM Transactions on Internet Technology (TOIT)*, 2(4):261–275, 2002.

[9] Michael D Gordon and Robert K Lindsay. Toward discovery support systems: A replication, re-examination, and extension of swanson's work on literature-based discovery of a connection

between raynaud's and fish oil. *Journal of the American Society for Information Science*, 47(2):116–128, 1996.

[10] Ruggero Gramatica, Tiziana Di Matteo, Stefano Giorgetti, Massimo Barbiani, Dorian Bevec, and Tomaso Aste. Graph theory enables drug repurposing–how a mathematical model can drive the discovery of hidden mechanisms of action. *PloS one*, 9(1):e84912, 2014.

[11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

[12] Dimitar Hristovski, Carol Friedman, Thomas C Rindflesch, and Borut Peterlin. Exploiting semantic relations for literature-based discovery. In *AMIA*, 2006.

[13] Dimitar Hristovski, Borut Peterlin, Joyce A Mitchell, and Susanne M Humphrey. Improving literature based discovery support by genetic knowledge integration. *Stud Health Technol Inform*, 2003.

[14] Dimitar Hristovski, Janez Stare, Borut Peterlin, and Saso Dzeroski. Supporting discovery in medicine by association rule mining in medline and umls. *Studies in health technology and informatics*, (2):1344–1348, 2001.

[15] Xiaohua Hu, Xiaodan Zhang, Illhoi Yoo, Xiaofeng Wang, and Jiali Feng. Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule. *International Journal of Intelligent Systems*, 25(2):207–223, 2010.

[16] Xiaohua Hu, Xiaodan Zhang, Illhoi Yoo, and Yanqing Zhang. A semantic approach for mining hidden links from complementary and non-interactive biomedical literature. In *SDM*, pages 200–209. SIAM, 2006.

[17] Andrej Kastrin, Thomas C Rindflesch, and Dimitar Hristovski. Large-scale structure of a network of co-occurring mesh terms: statistical analysis of macroscopic properties. *PloS one*, 9(7):e102188, 2014.

[18] Guangrong Li and Xiaodan Zhang. Mining biomedical knowledge using chi-square association rule. In *2010 IEEE International Conference on Granular Computing*, pages 283–285. IEEE, 2010.

[19] Robert K Lindsay and Michael D Gordon. Literature-based discovery by lexical statistics. *Journal of the Association for Information Science and Technology*, 50(7):574, 1999.

[20] James T MacClave and Terry Sincich. *Statistics*. Prentice Hall, 2003.

[21] Wanda Pratt and Meliha Yetisgen-Yildiz. Litlinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 105–112. ACM, 2003.

[22] NM Ramadan, H Halvorson, A Vande-Linde, Steven R Levine, JA Helpern, and KMA Welch. Low brain magnesium in migraine. *Headache: The Journal of Head and Face Pain*, 29(9):590–593, 1989.

[23] Neil R Smalheiser and Don R Swanson. Indomethacin and alzheimer's disease. *Neurology*, 46(2):583–583, 1996.

[24] Neil R Smalheiser and Don R Swanson. Linking estrogen to alzheimer's disease an informatics approach. *Neurology*, 47(3):809–810, 1996.

[25] Neil R Smalheiser and Don R Swanson. Calcium-independent phospholipase a2 and schizophrenia. *Archives of General Psychiatry*, 55(8):752–753, 1998.

[26] Padmini Srinivasan. Text mining: generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.

[27] Don R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.

[28] Don R Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, 1988.

[29] Don R Swanson, Neil R Smalheiser, and Abraham Bookstein. Information discovery from complementary literatures: categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology*, 52(10):797–812, 2001.

[30] Martin Theobald, Nigam Shah, and Jeff Shrager. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from pubmed guided by relationships in pharmgkb. *Summit on Translat Bioinforma*, 2009:124–128, 2009.

[31] Marc Weeber, Henny Klein, Alan R Aronson, James G Mork, LT De Jong-van Den Berg, and Rein Vos. Text-based discovery in biomedicine: the architecture of the dad-system. In *Proceedings of the AMIA Symposium*, page 903. American Medical Informatics Association, 2000.

[32] Marc Weeber, Rein Vos, Henny Klein, Alan R Aronson, Grietje Molema, et al. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association*, 10(3):252–259, 2003.

[33] Bartłomiej Wilkowski, Marcelo Fiszman, Christopher M Miller, Dimitar Hristovski, Sivaram Arabandi, Graciela Rosemblat, and Thomas C Rindflesch. Graph-based methods for discovery browsing with semantic predications. In *AMIA annual symposium proceedings*, volume 2011, page 1514. American Medical Informatics Association, 2011.

[34] Jonathan D Wren. Extending the mutual information measure to rank inferred literature relationships. *BMC bioinformatics*, 5(1):1, 2004.

[35] Tianyi Wu, Yuguo Chen, and Jiawei Han. Association mining in large databases: A re-examination of its measures. In *Knowledge Discovery in Databases: PKDD 2007*, pages 621–628. Springer, 2007.

[36] Meliha Yetisgen-Yildiz and Wanda Pratt. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of biomedical informatics*, 39(6):600–611, 2006.

[37] Meliha Yetisgen-Yildiz and Wanda Pratt. Evaluation of literature-based discovery systems. In *Literature-based discovery*, pages 101–113. Springer, 2008.