Clemson University
TigerPrints

All Dissertations | Dissertations

5-2012

Hybrid Query Expansion on Ontology Graph in Biomedical Information Retrieval

Liang Dong
Clemson University, ldong@g.clemson.edu

Follow this and additional works at: https://tigerprints.clemson.edu/all_dissertations

 Part of the Computer Sciences Commons

Recommended Citation
Dong, Liang, "Hybrid Query Expansion on Ontology Graph in Biomedical Information Retrieval" (2012). *All Dissertations*. 903.
https://tigerprints.clemson.edu/all_dissertations/903

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

HYBRID QUERY EXPANSION ON ONTOLOGY GRAPH IN BIOMEDICAL INFORMATION RETRIEVAL

A Dissertation
Presented to
The Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
School of Computing

By
Liang Dong
May 2012

Accepted by:
Dr. Zijun (James) Wang, Committee Chair
Dr. David P. Jacobs
Dr. Jim Martin
Dr. Pradip K. Srimani

ABSTRACT

Nowadays, biomedical researchers publish thousands of papers and journals every day. Searching through biomedical literature to keep up with the state of the art is a task of increasing difficulty for many individual researchers. The continuously increasing amount of biomedical text data has resulted in high demands for an efficient and effective *biomedical information retrieval* (BIR) system. Though many existing information retrieval techniques can be directly applied in BIR, BIR distinguishes itself in the extensive use of biomedical terms and abbreviations which present high ambiguity.

First of all, we studied a fundamental yet simpler problem of word semantic similarity. We proposed a novel semantic word similarity algorithm and related tools called Weighted Edge Similarity Tools (WEST). WEST was motivated by our discovery that humans are more sensitive to the semantic difference due to the categorization than that due to the generalization/specification. Unlike most existing methods which model the semantic similarity of words based on either the depth of their *Lowest Common Ancestor* (LCA) or the traversal distance of between the word pair in WordNet, WEST also considers the joint contribution of the weighted distance between two words and the weighted depth of their LCA in WordNet. Experiments show that weighted edge based word similarity method has achieved 83.5% accuracy to human judgments.

Query expansion problem can be viewed as selecting top *k* words which have the maximum accumulated similarity to a given word set. It has been proved as an effective method in BIR and has been studied for over two decades. However, most of the previous researches focus on only one controlled vocabulary: *MeSH*. In addition, early studies find that applying ontology won't necessarily improve searching performance. In this dissertation, we propose a novel graph based query expansion approach which is able to take advantage of the global information from multiple controlled vocabularies via building a biomedical ontology graph from selected vocabularies in *Metathesaurus*. We apply *Personalized PageRank* algorithm on the ontology graph to rank and identify top terms which are highly relevant to the original user query, yet not presented in that query. Those new terms are reordered by a weighted scheme to prioritize specialized concepts. We multiply a scaling factor to those final selected terms to prevent query drifting and append them to the original query in the search. Experiments show that our approach achieves 17.7% improvement in 11 points average precision and recall value against Lucene's default indexing and searching strategy and by 24.8% better against all the other strategies on average. Furthermore, we observe that expanding with specialized concepts rather than generalized concepts can substantially improve the recall-precision performance.

Furthermore, we have successfully applied WEST from the underlying WordNet graph to biomedical ontology graph constructed by multiple controlled vocabularies in Metathesaurus. Experiments indicate that WEST further improve the recall-precision performance.

Finally, we have developed a Graph-based Biomedical Search Engine (G-Bean) for retrieving and visualizing information from literature using our proposed query expansion algorithm. G-Bean accepts any medical related user query and processes them with expanded medical query to search for the MEDLINE database.

DEDICATION

I dedicate this dissertation to my grandmother, my parents, granduncle Yi Zhang, uncle Jin Pu's family, uncle Jun Zhang's family and uncle Pin Li's family, Aunt Wu Ping's family. I have my special thanks to my host family Ai Teh in Clemson. I couldn't have made this PhD degree without their love, support and encouragement during these five years.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. James Wang, for his precious advices and supports throughout my entire Ph.D. period.

I would also like to thank Dr. David P. Jacobs, Dr. Jim Martin and Dr. Pradip K. Srimani for serving on my committee and giving me advices on research during my Ph.D. study.

I would like to thank Lin Li, Alison Nolan, Yihua Ding and Xuebo Song for being great group mates.

I would like to thank Dr. Zhidian Du, Dr. Bo Li and Dr. Taylor William for many useful advices and suggestions.

Finally, I would like to thank the School of Computing and the Graduate School for providing me the wonderful learning experience.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

Chapter 1

Introduction

1.1. Problem Statement

Nowadays, biomedical researchers publish thousands of papers and journals every day. Searching through biomedical literature to keep up with the state of the art is a task of increasing difficulty for many individual researchers. The challenge is ever increasing in the scope of topical coverage as well as the fast-growing volume of biomedical literature [1, 2]. There is a high demand from the biological and medical community for an efficient and effective *biomedical information retrieval* (BIR) system. Though many existing information retrieval techniques can be directly used in BIR, BIR distinguishes itself in the extensive use of biomedical terminology as well as the high ambiguity those terms may present. One of the biggest challenges in BIR is to increase the recall and precision performance in searching *MEDLINE* database. MEDLINE [3] is the world's largest medical bibliographic database that contains more than 18.9 million citations (by July 2011) from approximately 5000 medical journals and articles. NCBI's *PubMed* [4]

system is the most widely used web interface for accessing MEDLINE, generally uses Boolean expressions to search the indexed documents.

However, effectively querying MEDLINE by PubMed is not an easy task for ordinary users. Due to the complexity of the query language for accurate searching result, the literature searching is usually performed by experienced search expert such as librarians [5]. It is widely reported [6, 7] that normal users, including those regularly use the PubMed system over the web, do not utilize the system as effectively as experts. Those inexperienced searchers either fail to employ the best query terms or fail to effectively apply Boolean expressions in the query statement [8]. In addition, since there is no one "correct" way to index an item, the disagreement between searchers and indexers under the Boolean systems can make inexperience searchers frustrated. One previous study [8] showed that the average novice searcher (third year medical student) requires 14 separate queries to attain their objective. In addition, users are often overwhelmed by the long list of search results: over one-third of PubMed queries result in 100 or more citations [2].

MEDLINE based information retrieval has been studied for more than two decades [9-11]. Those early studies observed that using controlled vocabularies such as MeSH offer no advantages in retrieval performance over free-text. The poor performance is caused by a number of potential reasons such as missing concepts and incomplete synonym sets [12].

Nevertheless, query expansion has been confirmed as an effective way to improve search performance. Srinivasan [13, 14] observed that *pseudo relevance feedback* (PRF)

based query expansion on MeSH vocabulary improved the retrieval performance. Yoo [15] and Abdou [16] re-designed the terms weight scheme found by PRF. However, since PubMed doesn't sort matched documents by relevance, the PRF strategy might not apply properly into PubMed.

There are two limitations for previous studies in query expansion: (1) only small amount of biomedical terms are used in indexing. Metathesaurus 2010AB covers total 2.3 million biomedical concepts, while most of the previous research only use MeSH along which only contains 26K terms in indexing. (2) Early studies did not consider the context information presented in the query. Expansion based on individual term may lead to the problem of query drifting.

Since the search mechanism in PubMed is not efficient for average users and existing methods have various drawbacks and limitations, a novel and better index and search approach is always desired in the biomedical community to overcome the shortcoming of the Boolean logic operation based PubMed system.

In recent years, we have continuously developed several original index strategies [17-21] in information retrieval and text mining and we applied them into MEDLINE [22, 23] based information retrieval and we have achieved great performance improvement over existing methods.

1.2. Dissertation Summary

This dissertation is dedicated to an original hybrid query expansion method in biomedical information retrieval by exploring ontology graph.

We first studied a relevant simple problem in natural language processing: *word semantic similarity problem*, which aims to compute the semantic similarity between two nodes in an ontology graph. We proposed a novel weighted edge word semantic similarity algorithm called WEST. We discovered an important human judgment difference between '*categorization*' pair and '*specification*' pair that humans are more sensitive to the semantic difference caused by the categorization than by specification. In other words, people view word pair separated by specification more similar than those separated by categorization. Base on this observation, we designed a set of strategies to measure word similarity considering that factor. Our proposed weighted edge distance model considers the specification level difference of a word pair and the specification level of its least common ancestor together. Based on this new model and a set of improved non-linear transfer functions, our method's result reaches a very good correlation against Miller-Charles's human similarity judgment.

The word semantic similarity gives us a hint that the similarity value exponentially decreases while the number of hops increases between two nodes. It also helps us abstract the query expansion problem into a mathematical model that we want to expand the user query with additional terms with the top accumulated similarity values, while preventing the problem of query drifting.

Our ontology graph exploration methodology applies personalized PageRank algorithm to the ontology graph. The original user query is used as the teleportation vector to compute a corresponding PageRank vector which is later used to construct the expanded query. As of our knowledge, this is the first personalized PageRank application

in text processing in biomedical information retrieval area. We hope this approach can bring interests and further studies from other researchers on personalized PageRank in biomedical information retrieval.

In addition, we applied this WEST word similarity algorithm from WordNet to multiple ontologies from Metathesaurus. The WEST algorithm is used to further filter the low similar personalized PageRank vector in order to provide screened expanded query.

Finally, we implement a web application of the biomedical search engine using our hybrid query expansion approach. The web application is open to the public and free to use, providing a better way for biomedical researchers to search for latest publications.

1.3. Research Contributions

New approach to query the MEDLINE database is always desirable in the biological and medical community. In this dissertation, we first studied a preliminary problem of word semantic similarity. Then, we extended the word semantic similarity into query expansion problem and proposed to apply Personalized PageRank to compute get the expansion candidates. We also apply the similarity algorithm to verify the confidence of these expanded terms.

Weighed Edge Word Semantic Similarity: first, we made an important observation that humans are more sensitive to the word semantic difference caused by the categorization than by specification. In another word, people view word pair separated by specification more similar than those separated by categorization. Our proposed weighted edge distance model merges the specification level difference of a word pair and the

specification level of its least common ancestor together. Based on this new model and a set of improved non-linear transfer functions, our method's result reaches a very good correlation against Miller-Charles's human similarity judgment.

Ontology Graph based Query Expansion: First of all, our proposed personalized PageRank based query expansion algorithm is conceptually novel and is very different from previous query expansion methods in information retrieval as of our knowledge. Unlike most of the previous ontology based studies which utilize only MeSH as their solo ontology, our personalized PageRank approach can employ multiple controlled vocabularies from Metathesaurus during the process. In this way, our system provides user with the ability to customize the underlying ontologies as they wish so that different user might be able to search the biomedical database using different underlying ontologies. For example, a biology scientist who is working on gene experiments can use the ontologies constructed by the single Gene Ontology (GO). To make the personalized PageRank algorithm work effectively, we have designed a systematic method to eliminate the mapped generalized biomedical concepts and populate closely related specialized concepts resulting in significant increase in the relevance of retrieval results. Our experimental analysis showed that eliminating generalized biomedical concepts in the search query may greatly improve the recall-precision performance. Finally, we demonstrate that query expansion based on ontology graph is more stable than that based on pseudo relevance feedback because sorting the retrieved documents by relevance is found to be often inaccurate.

Hybrid Approach: We have successfully explored and combined two different yet effective approaches to take advantages of the multiple biomedical ontologies into bioinformatics information retrieval. The final hybrid approach has further improved the performance of the search engine.

1.4. Dissertation Organization

The rest of the dissertation is organized as follows. In chapter 2 the background information of the ontology graph and biomedical information retrieval are presented. It also discusses existing query expansion methods, such as pseudo relevance feedback. In chapter 3, Weighted Edge Similarity Tools (WEST) is introduced to compute word semantic similarity on WordNet graph. The WEST method considers the difference of specification and generalization of a word pair in their positions in WordNet hierarchy. In chapter 4, the method and experimental results of query expansion using personalized PageRank algorithm is presented. In chapter 5, a hybrid query expansion algorithm is presented. The WEST algorithm is applied to the biomedical ontology graph and the expanded query from the personalized PageRank algorithm is further examined by the WEST algorithm to filter those concepts with low semantic similarity against the original query concepts. In chapter 6, a prototype of web application of the proposed query expansion biomedical information retrieval system is presented. Finally, conclusion and future work are presented in chapter 7.

Chapter 2

Background

2.1. Ontology

In philosophy, ontology is the study of being or existence and forms the basic subject matter of metaphysics. It seeks to describe the basic categories and relationships of being or existence to define entities and types of entities within its framework [24]. Ontology can be used to reason about the entities within that domain, and may be used to describe the domain. In computer science, an ontology represents an effective means of knowledge sharing within controlled and structured vocabulary [25]. Ontology provides a shared vocabulary, which can be used to model a domain — that is, the type of objects and/or concepts that exist, and their properties and relations. It is the structural framework for organizing information and is used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it. The creation of domain ontologies is also fundamental to the definition and use of an enterprise architecture framework [26].

In the following sections of this chapter, we are going to introduce several different ontologies for various purposes. First of all, WordNet [27] is a general English lexical ontology covering most of the common English concepts that supporting various purposes. In biomedical domain, the Metathesaurus of Unified Medical Language System (UMLS) framework [28, 29] includes many biomedical ontologies and terminologies such as Medical Subject Headings (MeSH) [30] and Medicine Clinical Term (SNOMED-CT) [31, 32]. NCBI Taxonomy [33] is another example of ontology to organize species where species in “is-a” relationships are grouped together using standard vocabulary.

2.2. WordNet

WordNet [27] is a lexical taxonomy database, widely used in many research fields such as artificial intelligence, natural language processing, information retrieval, and semantic web. WordNet provides a fine-grained structure ordering semantic word senses, called synsets, in a *Directed Acyclic Graph* (DAG). Senses/synsets of different “part-of-speech” are organized in different DAGs. All relationships form the edges in WordNet while the synsets, consist of the nodes in WordNet. Though WordNet 3.0 includes total 22 relationships between senses in its relationship hierarchy, the main relationship is still the “hypernym/hyponym (is-a)” relationship. The hypernym relationships of senses are shown in Figure 1 (only showing synset senses rather than synset ids for demonstration).

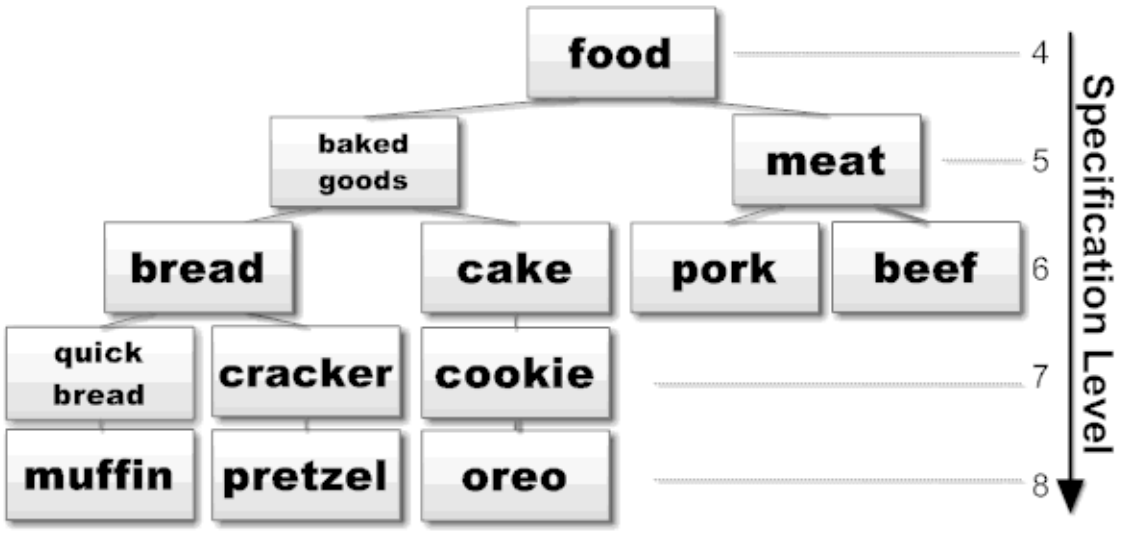


Figure 1: WordNet hierarchy

In WordNet, synsets and their relationships are used to model the polysemy and synonymy phenomena in English language. Polysemy means that one word has different meanings, while synonymy indicates different words represent the same concept/sense. If several words represent the same concept, it means they are synonymous and a single synset ID is assigned to them. For example, ‘lumber’ and ‘timber’ share the same concept, that is, “the wood of trees cut and prepared for use as building material”. Thus, these two words have the same synset ID in WordNet. As of the latest version 3.0 in 2006, the WordNet database contains 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs, and there are 101,863 monosemous and 60,384 polysemous noun words and senses.

Similarity of word senses obtained by WordNet-based methods closely matches the human perception because WordNet has coded the semantic relationships of word senses, as perceived by humans, into its hierarchical structure.

2.3. Medical Subject Headings

Medical Subject Headings (MeSH) [30], a subset of Unified Medical Language System (UMLS) [28, 29], is the U.S. National Library of Medicine's (NLM) controlled vocabulary thesaurus consisting of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. It is the main source vocabularies used with the primary purpose of supporting indexing, cataloging, and retrieval of medical literature articles stored in NLM MEDLINE database. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts and imposes uniformity and consistency in the indexing of biomedical literature. It is also used in the query-parser portion of PubMed's information retrieval system to map a user's query to MeSH descriptors in order to retrieve medical text that have been also indexed with the same MeSH descriptor.

There are three basic types of MeSH Records [34]: *Descriptors*, *Qualifiers*, and *Supplementary Concept Records* (SCRs). MeSH Descriptors, also known as Main Headings (MH), are used to index citations in NLM's MEDLINE database, for cataloging of publications, and other databases, and are reachable in PubMed as [MH]. Most Descriptors indicate the subject of an indexed item, such as a journal article, that is, what the article is about. Descriptors are generally updated on an annual basis but may, on occasion, be updated more frequently. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general levels of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders". More specific headings are found at more narrow levels of the twelve-level hierarchy, such as

“Ankle” and “Conduct Disorder”. There are 26,142 descriptors in 2011 MeSH, and over 177,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, “Vitamin C” is an entry term to “Ascorbic Acid”.

There are 83 different Qualifiers, also known as subheadings, used for indexing and cataloging in conjunction with Descriptors. Qualifiers afford a convenient means of grouping together those citations which are concerned with a particular aspect of a subject. For example, a “Liver/drug” effect indicates that the article or book is not about the “liver” in general, but about the effect of drugs on the “liver” Qualifiers are searchable in PubMed as MeSH Subheadings [SH]. Not all descriptor/qualifier combinations are allowed since some of them may be meaningless.

Supplementary Concept Records (SCRs) does not belong to the controlled vocabulary as such and are not used for indexing MEDLINE articles; instead they enlarge the thesaurus and contain links to the closest fitting descriptor to be used in a MEDLINE search. Many of these records describe chemical substances. SCRs are searchable by Substance Name [NM] in PubMed. Unlike Descriptors, SCRs do not have Tree Numbers; however, each SCR is linked to one or more Descriptors. SCRs are updated weekly, unlike Descriptor and Qualifier records, which are generally updated on an annual basis. There are currently over 199,000 SCR records within a separate thesaurus [3].

MeSH includes 16 high-level categories shown in the MeSH Tree Structure [35] where each category is assigned a letter: A for Anatomy, B for Organisms, C for Diseases, and so on. Each category is then repeatedly divided by a set of subcategories. When PubMed searches a MeSH term, it will automatically include narrower terms in the

search, if applicable. This is also called automatic explosion. Some terms occur in more than one place in the hierarchy. For example, “Eye” appears under the Anatomy branch, but also under the Sense Organs branch. Automatic explosion will include narrower terms from all instances of the term in the hierarchy.

2.4. Metathesaurus

The Metathesaurus [36] of Unified Medical Language System (UMLS) [28, 29] is a large, multi-purpose, and multi-lingual vocabulary database containing information about biomedical related concepts, their various names, and their inter-relationships.

The MeSH ontology we described in the previous section is also a part of the Metathesaurus ontology. Each biomedical concept is identified by a distinctive id called *Concept Unique Identifier* (CUI), which is an eight character alpha-numeric string. We use CUI to represent each biomedical concept in this dissertation. Each CUI is associated with a set of lexical variants strings, called *concept name*. The concept name may refer to medical conditions, appendages, diseases, drugs, and others; it may be single term, phrase, or a string of terms. Each concept is accompanied by an associated set of lexical variants cumulatively numbering over 1.7 million terms with 2 million strings representing a variation in concept spelling identified by a string identifier.

A depiction of concept organization as used in the Metathesaurus is shown in Figure 2. A concept is a grouping of synonymous terms; furthermore, each synonymous term listed for a concept contains acceptable spelling variations. These variations are

depicted as String 1 to String 4, while the synonymous terms are depicted as Term1 and Term 2.

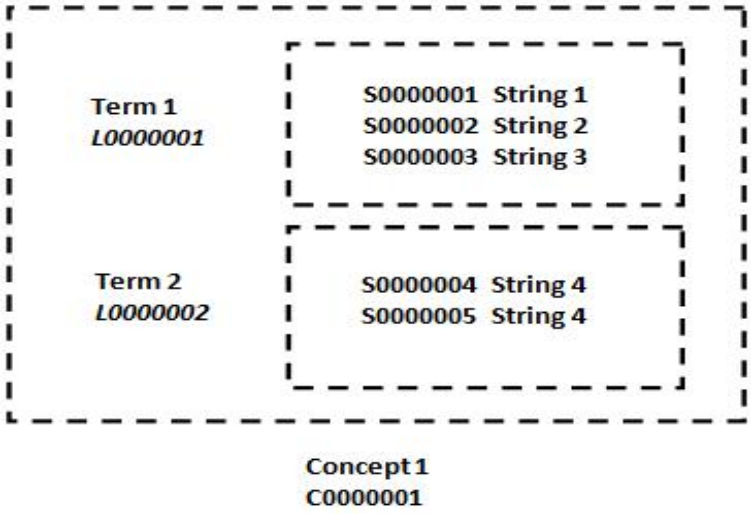


Figure 2: Metathesaurus concept organization

MRCON

```
C001403 | ENG | P | L0001403 | PF | S0010794 | Addison's Disease
C001403 | ENG | P | L0001403 | VC | S0352253 | ADDISON'S DISEASE
C001403 | ENG | P | L0001403 | VO | S0033587 | Disease, Addison
C001403 | ENG | P | L0001403 | VO | S0469271 | Addison's disease
C001403 | ENG | S | L0367999 | PF | S0469267 | Addison melanoderma
C001403 | ENG | S | L0373744 | PF | S0471237 | Asthenia pigmentosa
```

Figure 3: Metathesaurus `MRCONSO` table

The `MRCONSO` table in Figure 3 stores the entire CUIs and concept names. The `MRCONSO` table in consists of several data columns but the two of interests are concept name and CUI.

The Metathesaurus includes many inter-concept relationships as well. Most of these relationships come from individual vocabularies. The others are either added by NLM during Metathesaurus construction or contributed by users to support certain types

of applications. The inter-concept relationships are stored in the MRREL table depicted in Figure 4. Many types of relationships are included such as parent/child, immediate siblings.

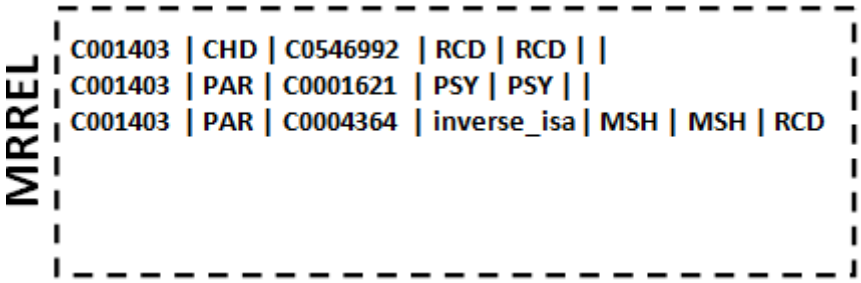


Figure 4: Metathesaurus MRREL table

2.5. Biomedical Information Retrieval

In computer science field of study, *information retrieval* (IR) [37] refers to finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). IR can also cover other kinds of data and information problems beyond that specified in the core definition above. The term “unstructured data” refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database, of the sort companies usually use to maintain product inventories and personnel records.

Nowadays, hundreds of millions of people engage in information retrieval every day when they use a web search engine such as Google or Bing. Information retrieval is fast becoming the dominant form of information access, overtaking traditional database-style searching.

The continuously increasing amount of biomedical information has resulted in higher demands for an efficient and effective *biomedical information retrieval* (BIR) system. This requires the ability to systematically compare large data sets with all the knowledge that is derived from the published data, which allows the biological relevance of the data set to be interpreted. The information, which is measured in terms of the numbers of articles and journals that are published, is increasing at a considerable rate, so that it is no longer possible for a researcher to keep up to date with all the relevant literature manually, even on specialized topics.

Figure 5 shows the numbers of journals, papers (as represented by MEDLINE abstracts), papers on the cell cycle and papers on Cdc28 that were published each year from 1950 to 2005 [1]. An average for 3 years was calculated for the Cdc28 curve because of much lower numbers. The number of new papers that were published each year continues to increase, especially on certain topics such as the cell cycle, for which it is no longer possible to read all new papers that are published. By contrast, specific proteins that are “hot” at one point in time tend to lose their popularity later, as exemplified by Cdc28.

Though many existing information retrieval techniques can be directly used in biomedical information retrieval, BIR distinguishes itself in the extensive use of biomedical terminology which contains many uncommon terms and ambiguous abbreviations..

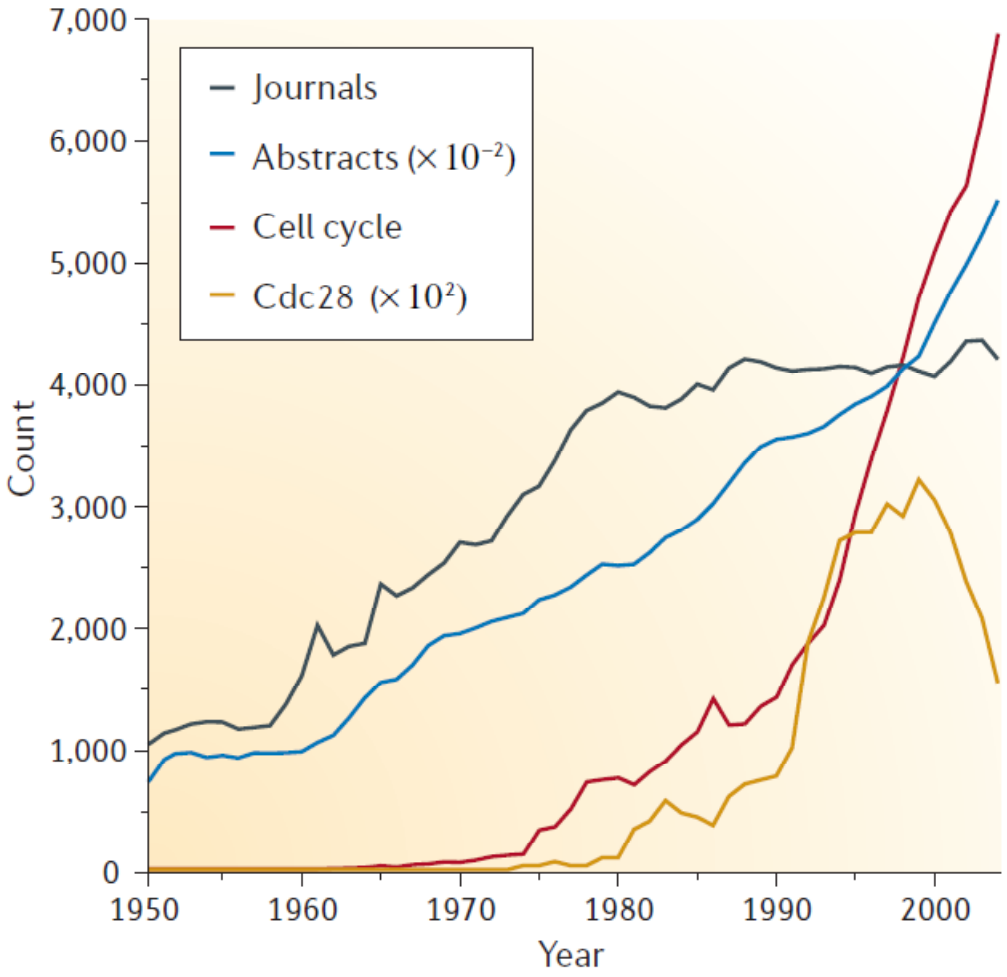


Figure 5: Increasing trend of publications containing gene “Cdc28”
(cited from Jenson 2005 [1])

2.6. MEDLINE and PubMed database

Advances in biotechnology, together with the widespread use of high-throughput methods for gene analysis, have helped shifting the focus of biological research from specific genes and proteins to a more systemic analysis of the underlying biological problem. Researchers now face the increasing need to plan their experiments and analyze

the resulting datasets in view of the quickly expanding biomedical information available [38].

Medical Literature Analysis and Retrieval System Online (MEDLINE) [3] is the National Library of Medicine's premier database that hosts medical journals and articles in the life sciences with a concentration in biomedicine. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution [39]. It also leverages a controlled vocabulary, meaning that there is a specific set of terms used to describe each stored article; describing each article is generally known as indexing. Records in MEDLINE are indexed with the MeSH vocabulary to facilitate retrieval by regular users, researchers, students, and doctors. Users who are familiar with the MeSH vocabulary are typically better searchers then those users who are unfamiliar with the specialized vocabulary. The records in MEDLINE are covered from 1946 to present, with some even older materials.

PubMed [4], as the most popular biomedical information retrieval system, gives researchers access to over 17 million citations from a broad collection of scientific journals, indexed by the MEDLINE literature database. PubMed is a web-based information retrieval system developed by the National Center for Biotechnology Information (NCBI) to provide access to citations from biomedical literature. PubMed facilitates access to the biomedical literature by combining the MeSH based indexing from MEDLINE, with Boolean and vector space models for document retrieval, offering

a single interface from which these journals can be searched [40]. The result of a MEDLINE/PubMed search is a list of citations (including authors, title, source, and often an abstract) to journal articles and an indication of free electronic full-text availability. Searching is free of charge and does not require registration. Searching MEDLINE/PubMed effectively is a learned skill; untrained users are sometimes frustrated with the large numbers of articles returned by simple searches.

The weaknesses of the PubMed information retrieval system are made manifest when indexing medical articles and resolving users search queries to indexes. In an effort to build an information retrieval system based on semantic retrieval, PubMed has heavily utilized the MeSH vocabulary in its indexing and user-querying components. There are 26,142 descriptors, 83 qualifiers, over 177K assisting entry terms and over 199K supplementary concept records in MeSH 2011; but only descriptors and qualifiers are used in indexing MEDLINE. In comparison, NLM *Metathesaurus* 2010AB covers 2.3 million biomedical concepts. The primary disadvantage of the MEDLINE/PubMed system is that it indexes millions of documents with less than 1.1% of the available biomedical vocabulary. This disadvantage is obvious when retrieving results from PubMed that are semantically close to the information requested, but not sufficiently narrow resulting in very low precision and recall and requiring multiple searches by users.

2.7. Query Expansion

Previous sections introduced biomedical related information retrieval. The next two sections discuss the related techniques we will use in this dissertation.

Query Expansion (QE) is the process of reformulating an original query to improve retrieval performance in information retrieval. In the context of web search engine, query expansion involves evaluating a user's input (what words were typed into the search query area and sometimes other types of data) and expanding the search query to match additional documents. Search engines invoke query expansion to increase the quality of user search results assuming that users do not always formulate search queries using the best terms [41].

The goal of query expansion is to increase recall, but precision can potentially increase as well, by including those records which are more relevant or at least equally relevant into the query result set. Those records which have the potential to be more relevant to the user's desired query would be included by applying query expansion. At the same time, many of the current commercial search engines use Term Frequency – Inverse Document Frequency (TF-IDF) to assist in ranking. By ranking the occurrences of user's input as well as synonyms and alternate morphological forms, documents with a higher density (high frequency and close proximity) tend to migrate higher up in the search results, leading to a higher quality of the search results near the top of the results, despite the larger recall.

Query expansion techniques can broadly be classified into three categories:

(1) *Collection based or global analysis*: use context global of terms in collection to find out similar terms with query terms [42].

(2) *Query based or local analysis*: the context of terms is reduced to smaller subsets of information which is given from relevance feedback or pseudo relevance feedback [43] and collaboration information like user profile, query logs [44].

(3) *Knowledge based approach*: the exploration of the knowledge in external knowledge sources, mostly with general domain thesaurus like WordNet. They explore semantic links in the ontology graph in order to find out in the related terms of query concepts to expand.

In this dissertation, our proposed query expansion approach is knowledge based approach.

2.8. Pseudo Relevance Feedback

In information retrieval systems, *relevance feedback* (RF) is an effective query expansion technique. It takes the results that are initially returned from a given query and it relies on user interaction to identify the relevant results to build and perform a new query.

Pseudo Relevance Feedback (PRF) automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction. The method performs normal retrieval to find an initial set of most relevant

documents; it then assumes that the top “k” ranked documents are relevant; and it finally performs relevance feedback as before under this assumption [45].

The success of relevance feedback depends on certain assumptions [37]: Firstly, the user has to have sufficient knowledge to be able to make an initial query which is at least somewhere close to the documents they desire. This is needed anyhow for successful information retrieval in the basic case, but it is important to see the kinds of problems that relevance feedback cannot solve alone. Cases where relevance feedback alone is not sufficient include:

• Misspellings. If the user spells a term in a different way to the way it is spelled in any document in the collection, then relevance feedback is unlikely to be effective. This can be addressed by the spelling correction techniques.

• Cross-language information retrieval. Documents in another language are not nearby in a vector space based on term distribution. Rather, documents in the same language cluster more closely together.

• Mismatch of searcher’s vocabulary versus collection vocabulary. If the user searches for laptop but all the documents use the term notebook computer, then the query will fail, and relevance feedback is again most likely ineffective.

Secondly, the relevance feedback approach requires relevant documents to be similar to each other. That is, they should cluster. Ideally, the term distribution in all relevant documents will be similar to that in the documents marked by the users, while the term distribution in all non-relevant documents will be different from those in relevant documents. Things will work well if all relevant documents are tightly clustered

around a single prototype, or, at least, if there are different prototypes, if the relevant documents have significant vocabulary overlap, while similarities between relevant and non-relevant documents are small. Implicitly, the Rocchio relevance feedback model treats relevant documents as a single *cluster*, which it models via the centroid of the cluster. This approach does not work as well if the relevant documents are a multimodal class, that is, they consist of several clusters of documents within the vector space. This can happen with:

• Subsets of the documents using different vocabulary, such as Burma vs. Myanmar

• A query for which the answer set is inherently disjunctive, such as Pop stars who once worked at Burger King.

• Instances of a general concept, which often appear as a disjunction of more specific concepts, for example, felines.

Chapter 3

Weighted Edge Similarity Algorithm and Tools

3.1. Motivation

Determining the semantic similarity of two words is useful yet challenge. The measure of the semantic similarity of words is a building block in many important applications, such as word sense disambiguation, clustering, embedding, ranking, and spell-checking. However, polysemy and synonymy phenomena widely exist in natural language, and psychologists have demonstrated that the human perception of the similarity between words is subject to the context. Therefore, it is extremely difficult to model the human perspective on the semantic similarity of words.

In recent two decades, researchers have tried to solve this hard problem through different approaches. Existing methods can be divided into two categories:

Thesaurus-based methods rely on a human-built thesaurus, such as WordNet. Wu and Palmer[46] consider the specification level of two word senses and their least common ancestor, but their linear similarity function is simple, which is not accurate with human judgments. Li et al. [47] proposed an efficient non-linear method and achieved

significant performance improvement over other studies [48-51]. *Information content*,[49-52], statistical word distribution of text corpus, is used as supplemented information. Several corpuses, including *Brown corpus, Semcor,* and *Treebank*, are used to acquire the information content. However, if two words are well annotated near the root of the thesaurus, called *shallow annotation*, their semantic distance will always be computed close to zero, thus causing abnormal high similarity result.

Knowledge-based Methods take advantage of human knowledge base. Cilibrasi et al. [53] proposed *Normalized Google Distance*, which assumes that the semantic similarity of two words is associated to the number of web pages returned by Google search engine. However, Normalized Google Distance only reflects the concurrency in textural document. It is not really a concept distance since it doesn't preserve triangle property ESA [54] maps each word into a vector of a set of articles derived from Wikipedia corpus by traditional Vector Space Model. Then, relatedness is measured by the cosine of the angle of two Wikipedia-article vectors. Personalized PageRank [55] is used on WordNet graph.

To address the drawbacks of these existing methods, we propose WEST, a new method to consider the co-locations of word pairs with their *Least Common Ancestor* (LCA): when two different word pairs that share the same LCA and have the same graph distance, the similarity value of one word-pair should not always be the same of the other. Actually, it should be decided by the specification levels of each individual word. The advantages of this new method are two folds: (1) the semantic similarity of words measured by this method closely matches the human perspective; (2) the measure of the

semantic similarity relies only on co-location information of words within the WordNet, thus more computation effective than those requiring the computation of corpus statistics. Experimental studies show that our proposed method outperforms all existing methods.

The rest of this chapter is organized as follows. We introduce the background knowledge of word similarity in section 3.2. In section 3.3, we observe the difference between word pair's inheritance and categorization relations. Then, we propose the weighted edge method to model the semantic distance of words in section 3.4. We discuss the benchmark, dataset, methods of the experimental studies. We discuss experimental result in section 3.5. Section 3.6 shows the architecture and implementation of WEST -- a set of web tools for public use. Finally, we have our conclusion in section 3.7.

3.2. Semantic Similarity of Words

Many recent studies have employed WordNet as their knowledge base to study the semantic relationships between words. WordNet [56] is a lexical taxonomy database, widely used in many research fields such as natural language processing, data mining, and information retrieval. It provides a fine-grained structure ordering semantic word senses or *synset*s, in a *Directed Acyclic Graph* (DAG) hierarchy, as shown in Figure 6. Words of different "part-of-speech" are organized in different DAGs. Although WordNet 3.0 includes total 22 relationships between words in its relationship hierarchy, the main relationship is still the "hypernym/hyponym" inheritance relationship. All relationships form the edges in WordNet, and the word senses, or synsets, consist of the nodes in

WordNet. Synsets and their relationships are used to model the polysemy and synonymy phenomena in English language.

Polysemy means that one word has different meanings, while synonymy indicates different words represent the same concept. The statistics show that there are 101,863 monosemous and 60,384 polysemous noun words and senses in WordNet 3.0. If several words represent the same concept, it means they are synonymous and a single synset ID is assigned to them. For example, ‘lumber’ and ‘timber’ share the same concept, that is, “the wood of trees cut and prepared for use as building material”. Thus, these two words have the same synset ID in WordNet.

Previous studies [47] have identified two critical factors influencing semantic similarity: *graph distance*, and s*pecification level* (SpecLev) of their *Least Common Ancestor* (LCA). Graph distance counts the number of hops on the shortest path between two synsets, and specification level (SpecLev) is the number of hops on the shortest path from the synset to its root, or the *depth* of synset in WordNet. If a synset is closer to the root in the WordNet, it has a lower SpecLev, thus has a more general meaning.

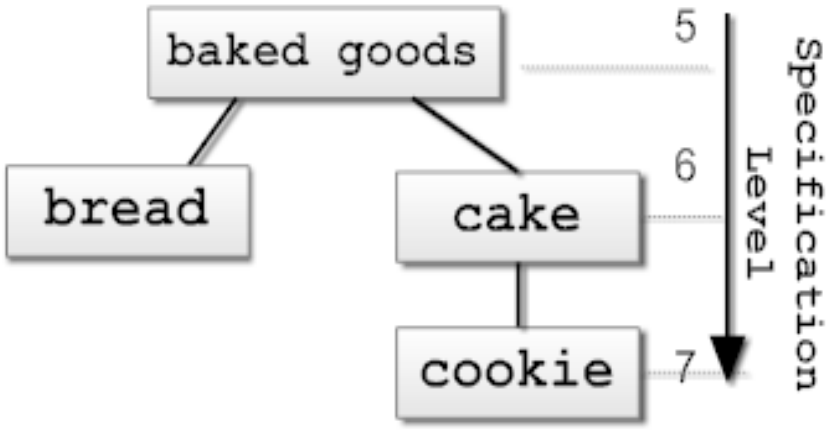


Figure 6: WordNet specification level

3.3. Inheritance vs. Categorization

If only the graph distance and the SpecLev of their LCA are used to measure the semantic similarity of two words, when two different word pairs share the same LCA and the graph distance between the words in one pair is the same as that in another pair, the semantic similarity of one word pair is measured to be the same as the semantic similarity of another word pair using existing methods. Does this match the human perspective? None of the existing studies have investigated this issue.

To study how human beings judge the semantic similarity of words in the aforementioned situations, we select two word-pairs that share the same LCA and the words within each pair have the same graph distance. In one word-pair, called *categorization pair*, the words are both descendants of their LCA, since they are separated into different categories. In another word pair, called *inheritance pair*, one word is descendant of another word. We put these two word-pairs together as a comparison group. In Figure 6, “bread-cake” is a categorization pair; “baked goods-cookie” is an inheritance pair. These two pairs have the same LCA “baked goods”, and the graph distance of “bread-cake” and “baked goods-cookie” are 2.

We collect 20 groups of such comparison pairs. The graph distance of the word-pairs in the first 10 groups is 2 in Table 1. The graph distance of the word-pairs in the second 10 groups, shown in Table 2, is 4. Then we randomly stop people in Clemson University campus and ask them to judge which pair in each comparison group is more similar semantically. 51 individuals finished the questionnaire anonymously. In Table 1 and Table 2, each row contains a group of word-pairs. The left is the inheritance pair and

the right is the categorization pair. The number in the second column represents the number of people who think the inheritance pair is more similar semantically. The number in the last column represents the number of people who feel the categorization pair is more similar. For those who feel both pairs are semantically equal or who cannot tell which pair is more similar, no number is added to any column. The survey results in Table 1 shows that in 68.41% of cases of graph distance at 2, people think the inheritance pairs are more similar, and in 31.59% of cases, people think the categorization pairs are more similar. The results in Table 2 demonstrate that in 76.67% of cases of graph distance at 4, people think that the inheritance pairs are more similar, and in 23.33% vice versa.

Table 1: Comparison groups with graph distance equal to 2 in each pair

Inheritance Word-Pair			***Categorization Word-Pair***	
baked-goods :: cookie	30	⇔	bread :: cake	19
beef :: food	48	⇔	meat :: chocolate	2
brownie :: cake	44	⇔	cookie :: fruitcake	5
ground beef :: meat	24	⇔	pork :: mutton	25
apple pie :: pastry	42	⇔	pie :: puff	8
stove :: device	41	⇔	comb :: fan	8
engine :: machine	18	⇔	computer :: calculator	33
hunting dog:: canine	27	⇔	wolf :: fox	22
minicab :: car	29	⇔	jeep :: sedan	21
gold :: metal	37	⇔	aluminum :: zinc	14
Total	**340**			**157**

Table 2: Comparison groups with graph distance equal to 4 in each pair

Inheritance Word-Pair			***Categorization Word-Pair***	
apple pie :: food	44	⇔	cake :: beef	3
clementine :: fruit	36	⇔	apple :: almond	15
chicken :: food	47	⇔	octopus :: pastry	0
dynamo :: machine	45	⇔	engine :: abacus	4
abbey :: building	26	⇔	hostel :: mansion	23
tabloid :: medium	8	⇔	broadcasting :: journalism	43
laptop :: computer	51	⇔	workstation :: chatroom	0
American football :: athletic game	36	⇔	golf :: basketball	14
cliff diving :: sports	44	⇔	hunting :: swimming	6
collegiate dictionary :: book	41	⇔	atlas :: bestseller	7
Total	**378**			**115**

Our survey results have revealed an interesting observation that people are more sensitive to the semantic difference caused by categorization than by the inheritance/specification. They think two words in different categories are less similar than two words separated only by specification levels when graph distances are the same. This is more obvious when the graph distance of the words becomes longer. This important fact has never been discovered in any previous studies. To reflect the true human perception in measuring the semantic similarity of words, we have to include this critical factor in our measurement model. We define *Specification Level Difference* (SLD) as absolute difference of the SpecLev of two word senses. Given two word senses (ws_i, ws_j), $slev_i, slev_j$ are their corresponding SpecLev, the SLD is measured as Equation (1):

$$SLD(ws_i, ws_j) = | slev_i - slev_j | \quad (1)$$

SLD models the impact factor of the inheritance and the categorization on the semantic similarity of two synsets with the same graph distance and the same LCA.

3.4. Our Weighted Edge Semantic Similarity Approach

3.4.1. Weighted Edge

Since WordNet architecture is ordered by word sense, we assume the semantic similarity of a word-pair is the highest semantic similarity value measured from all its sense-pairs. The semantic similarity of a senses pair (ws_i, ws_j) can be determined by three factors in the WordNet Hierarchy:

*(**Factor 1**) Specification Level of its LCA $slev_{lca}$ on the shortest path linking the sense-pair;*

*(**Factor 2**) The shortest graph distance $l_{gd}(ws_i, ws_j)$ between the sense-pair;*

*(**Factor 3**) Specification Level Difference $SLD(ws_i, ws_j)$ between the sense-pair.*

An intuitive approach to measure the semantic similarity of a sense pair is to summarize these three factors under proper scaling parameters. However, it is very hard to determine three proper scaling parameters due to their correlations. In this section, we propose a simple yet effective method to measure the semantic similarity of sense pair based on the combined effect of these factors.

We propose a simple yet effective method to measure the semantic similarity of sense pair based on the previous observation.

Given a word pair, we query WordNet for all its sense pairs to find which LCA on the path has the highest SpecLev, since SpecLev of LCA is the most decisive factor in similarity measurement. If more than one sense pairs are found, the sense pair with the shortest graph distance in WordNet is selected. Then, we can focus on measuring the similarity of corresponding sense-pair. This process is similar to the "disjunctive concepts" method by Rada [48] and Resnik [49] respectively, but the difference in our method is that, during the sense-pair selection, we consider the SpecLev of LCA in the first place rather than graph distance in previous studies. This adjustment is based on the observation that the SpecLev of LCA plays the most vital role.

Given a synset pair (ws_i, ws_j) with SpecLev $(slev_i, slev_j)$, and the SpecLev of their LCA (ws_{lca}) be $slev_{lca}$, we can represent the graph distance $l_{gd}(ws_i, ws_j)$ between ws_i and ws_j as the sum of $SLD(ws_i, ws_{lca})$ and $SLD(ws_j, ws_{lca})$ in Equation (2):

$$\begin{aligned} & l_{gd}(ws_i, ws_j) \\ & = SLD(ws_i, ws_{lca}) + SLD(ws_j, ws_{lca}) \\ & = | slev_i - slev_{lca} | + | slev_j - slev_{lca} | \\ & = slev_i + slev_j - 2 \cdot slev_{lca} \end{aligned} \tag{2}$$

We assume each edge in the WordNet hierarchy has a weighted value, which is an exponential decreasing value associated to its SpecLev. A coefficient $\alpha \in (0,1]$ is used to represent the *weight decreasing rate* α along the edge of WordNet.

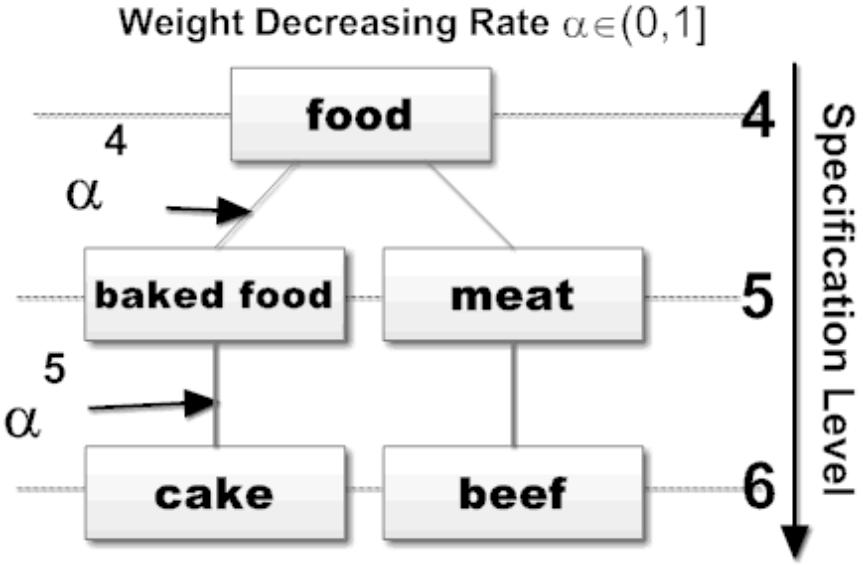


Figure 7: Weighted Edge Decreases along its SpecLev

As shown in Figure 7, we define the edge weight e_k be $e_k = \alpha^k$ for the edge connecting two neighboring synsets at level k and $k+1$ respectively. Thus, the edge linking the root node $(k=0)$ and first level nodes $(k=1)$ has an edge weight $\alpha^0 = 1$. The more specific or deeper an edge locates in the WordNet hierarchy, the smaller weight it has.

Using our weighted edge model, we define *weighted edge distance*(l_w) between a sense pair (ws_i, ws_j), as a function f of the three SpecLev values $(slev_{lca}, slev_i, slev_j)$. That is,

$$\begin{aligned} l_w &= f\,(slev_{lca}, slev_i, slev_j) \\ &= \sum_{m=slev_{lca}}^{slev_i - 1} e_m + \sum_{n=slev_{lca}}^{slev_j - 1} e_n \end{aligned} \tag{3}$$

Weighted edge distance is the sum of all the edge weights along its shortest path to its LCA. Given a weight decreasing rate $\alpha \in (0,1]$, we substitute e_k with a^k in Equation (3), we have

$$\begin{aligned} l_w &= f(\alpha, slev_{lca}, slev_i, slev_j) \\ &= \alpha^{slev_{lca}} \cdot \left(\sum_{m=0}^{slev_i - slev_{lca} - 1} \alpha^m + \sum_{n=0}^{slev_j - slev_{lca} - 1} \alpha^n \right) \end{aligned} \tag{4}$$

Our approach generalizes the traditional graph distance. When $\alpha = 1$, the weighted edge distance turns into the traditional graph distance. When $\alpha \in (0,1)$, the edge value exponentially decreases with the increase of SpecLev along the hierarchy.

We can pre-compute weighted edge distance for each SpecLev to its root (SpecLev 0) to accelerate the computation for any sense pair in constant time. The measurement of Weighted Edge Distance l_w for sense pair (ws_i, ws_j) with LCA (ws_{lca}) can be optimized as:

$$\begin{aligned} & l_w(ws_i, ws_j) \\ &= l_w(ws_i, ws_{root}) + l_w(ws_j, ws_{root}) - 2 \cdot l_w(ws_{lca}, ws_{root}) \end{aligned} \tag{5}$$

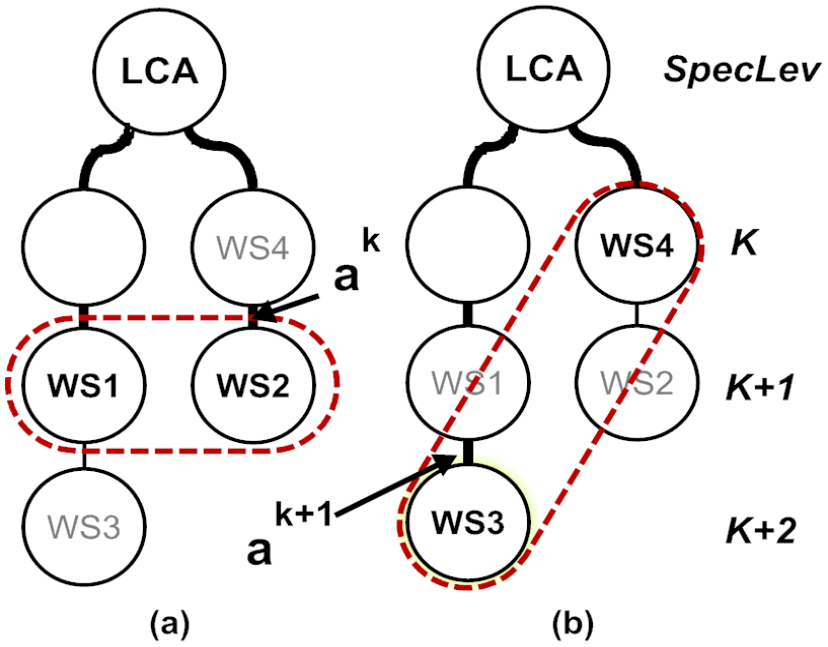


Figure 8: Increasing Specification Level Difference from 0 in (a) to 2 in (b)

Next, we'll show how our weighted edge model and the proposed Equation (6) reflect the human perception difference between inheritance and categorization. As illustrated in Figure 8, given two sense-pairs (ws_1, ws_2) and (ws_3, ws_4), which have the same graph distance and share the same LCA, but the SpecLev difference of (ws_1, ws_2) is zero and the SpecLev difference of (ws_3, ws_4) is two. According to definition of Weighted Edge Distance, we have:

$$\begin{aligned} & l_w(ws_1, ws_2) - l_w(ws_3, ws_4) \\ & = \alpha^k - \alpha^{k+1} = \alpha^k(1-\alpha) \geq 0, \alpha \in (0,1] \end{aligned} \tag{6}$$

Thus, $l_w(ws_1, ws_2) \geq l_w(ws_3, ws_4)$ denotes sense-pair (ws_3, ws_4) is more similar than (ws_1, ws_2) which is coherent with human judgments. We can conclude that given a sense-pair with a fixed graph distance, the increase of their Specification Level Difference from Figure 8 (a) to (b) reduces its Weighted Edge Distance, meaning the

sense-pair has a higher similar value. This result conforms to our discovery that humans are more sensitive to the semantic difference caused by categorization than that caused by specification/inheritance.

3.4.2. New Transfer Function

Now, we need to design a transfer function g to convert the Weighted Edge Distance to semantic similarity value. We define the semantic similarity between sense pair ws_i and ws_j or $sim(ws_i, ws_j)$ be a function of its weighted edge distance l_w:

$$sim(ws_i, ws_j) = g(l_w) \tag{7}$$

To efficiently calculate $sim(ws_i, ws_j)$, an approximation function that should demonstrate the following three features:

(1) It should be a continuous function with variable range $[0, +\infty)$ and value range [0,1];

(2) When the Weighted Edge Distance is 0, the similarity value should be 1. It means the two word sense share the same synset/concept;

(3) When the Weighted Edge Distance approaches the positive infinite, the similarity should be 0, meaning the two words are far away with each other conceptually.

Li's method [47] used both linear and non-linear functions to approximate the traditional graph distance to the similarity value between 0 to 1. His result showed that the non-linear function performs remarkably better than linear function. We further

extend his experimental studies with six different non-linear functions and found that the similarity values obtained by hyperbolic functions best match human judgments.

Two hyperbolic functions are used as our approximate functions. One is *Hyperbolic Secant* (Sech) and the other is *Hyperbolic Tangent Cardinal* (Tanhc). Both hyperbolic functions are monotonically decreasing functions of x with the value range from 0 to 1.

3.5. Validation of Weighted Edge Similarity Approach

3.5.1. Benchmark Datasets

It is ideal that the semantic similarity of words measured by our method matches perfectly with the human perception. Therefore, it is reasonable to compare the semantic similarity values obtained by our method with human judgments. Correlating the computed semantic similarity measures with human judgments is a common practice in evaluating the similarity measurement techniques.

In 1965, German scientists Rubenstein and Goodenough [57] presented 51 human subjects with 65 noun pairs (called RG set) and asked them to scale the similarity from 0.0 to 4.0 for “no similarity” to “perfect synonymy”. 25 years later, Miller and Charles [38] in USA divided the RG Set into three semantic similar parts with high, medium, and low similar level. They choose 10 word pairs from each level and repeated the Rubenstein-Goodenough procedures with 38 undergraduate students. The 30 word pairs are named Miller-Charles (MC) set. It is worth noting that the correlation between the

two experimental results is as high as 0.97, indicating that human judgment is quite stable under little influence from time span and language difference. Again, Resnik [49] replicated the same experiment on the MC set, presenting them to 10 graduate students or postdoctoral researchers at the University of Pennsylvania. The correlation between Resnik rating and Miller-Charles rating was 0.96, quite close to the 0.97 correlation in the earlier study. Resnik computed average correlation between individual subjects' rating with MC rating to be 0.88, with a standard deviation of 0.08. He claimed the correlation value 0.88 represents an upper bound from a computational attempt to perform the same task.

Many previous studies [49-51] used Miller and Charles [38] MC set as the comparison baseline. Since the earlier version of WordNet missed word "woodland" from the MC set, only 28 word pairs were used in these studies. Li et al. utilized all 65 pairs of the original Rubenstein-Goodenough set. Since MC set is a subset of RG set, Li applied the 28 pairs of MC set as testing set D_0, and the rest 37 pairs of words as training set D_1. He tried ten different strategies, obtained the optimal parameters on training set D_1 and evaluated the performance of his strategies on testing set D_0 dataset.

In this section, we conduct similar experiments using our proposed scheme on different strategies and calculate the correlation between our computed similarities and the human judgments. Due to Li's method [47] being regarded as "particularly effective, best and fastest" according to Varelas [58], we also repeated Li's experiments with his best strategy, using the same training set and testing set respectively. As in Li's study, we obtain the optimal parameter values using the training set D_1, then we run the testing set

D_0 with these optimal parametric value. Finally, we compare the experimental results obtained by our method with those by Li's method.

We list the complete information of testing dataset D_0 as well as training dataset D_1 in Table 3 and Table 4 respectively.

Table 3: Testing data of MC dataset

Word1	Word2	Graph Distance	SpecLev LCA	SpecLev Word1	SpecLev Word2
cord	smile	10	1	6	6
rooster	voyage	23	0	13	10
noon	string	11	1	9	4
glass	magician	9	3	7	8
monk	slave	4	6	9	7
coast	forest	5	2	5	4
monk	oracle	7	6	9	10
lad	wizard	4	6	8	8
forest	graveyard	8	2	4	8
food	rooster	15	1	4	13
coast	hill	4	3	5	5
car	journey	18	0	9	9
crane	implement	4	5	8	6
brother	lad	4	6	8	8
bird	crane	3	9	9	12
bird	cock	1	9	9	10
food	fruit	9	2	5	8
brother	monk	1	9	10	9
asylum	madhouse	1	9	9	10
furnace	stove	9	4	9	8
magician	wizard	0	8	8	8
journey	voyage	1	9	9	10
coast	shore	1	4	5	4
implement	tool	1	6	6	7
boy	lad	1	8	8	9
automobile	car	0	11	11	11
midday	noon	0	9	9	9
gem	jewel	0	8	8	8

Table 4: Training data of MC dataset

Word1	Word2	Graph Distance	SpecLev LCA	SpecLev Word1	SpecLev Word2
autograph	shore	9	0	5	4
automobile	wizard	12	3	11	8
mound	stove	7	5	9	8
grin	implement	12	0	6	6
asylum	fruit	6	4	7	7
asylum	monk	10	3	7	9
graveyard	madhouse	14	2	8	10
boy	rooster	11	5	8	13
cushion	jewel	6	4	6	8
asylum	cemetery	11	2	7	8
grin	lad	11	0	6	8
shore	woodland	4	2	4	4
boy	sage	5	6	8	9
automobile	cushion	8	5	11	8
mound	shore	9	5	9	10
cemetery	woodland	8	2	8	4
shore	voyage	14	0	4	10
bird	woodland	9	2	9	4
furnace	implement	7	4	9	6
crane	rooster	7	9	12	13
hill	woodland	5	2	5	4
cemetery	mound	10	2	8	6
glass	jewel	7	4	7	8
magician	oracle	6	6	8	10
sage	wizard	5	6	9	8
oracle	sage	5	7	10	9
hill	mound	0	9	9	9
cord	string	1	6	6	7
glass	tumbler	1	7	7	8
grin	smile	0	6	6	6
serf	slave	3	7	10	7
autograph	signature	1	6	7	6
forest	woodland	0	4	4	4
cock	rooster	0	13	13	13
cushion	pillow	1	6	6	7
cemetery	graveyard	0	8	8	8

3.5.2. Experiments on Different Strategies

We propose eight different strategies to calculate semantic similarity in this section. The first two strategies is replication of Li's 3rd and 4th Strategy for comparison purpose, then we conduct six new strategies combining weight edge distance and new transfer functions.

Li's strategy uses graph distance (l_{gd}) and SpecLev of their LCA ($slev_{lca}$) to calculate the similarity value between two synsets. In our strategies, only the Weighted Edge Distance l_w is used to calculate the semantic similarity of words.

To ensure the computed similarities obtained by transfer function matches with human judgments as closely as possible, we need to find an optimal *Weighted Decreasing Rate* α. For each strategy, we use the training set D_1 to obtain the optimal α value. We vary α from 0.05 to 1 with an increment of 0.05, and calculate the correlation between the computed similarities and human judgments on training set D_1. The α value that yields the highest correlation between computed similarities and human judgments is selected as the optimal parameter.

We note that Li's second transfer function requires two tuning factors (α, β). For this function, we vary α from 0.05 to 1 with an increment of 0.05 and β between 0.1 and 1 with an increment of 0.1. Values of α and β yielding the highest correlation between computed similarities and human judgments will be selected as the optimal parameters. The optimal parameters obtained by training set D_1 will then be used to calculate the semantic similarity values for word-pairs in testing set D_0. Finally, $\alpha = 0.20$

we calculate the correlations between the computed similarity values and Miller and Charles's human judgments on these word-pairs.

Strategy 1: We repeat Li et al. [47]'s Strategy 3rd as our first strategy. Li showed that non-linear function greatly improves the semantic similarity measure. A monotonically decreasing function $g_1(x) = e^{-x}$ is used to approximate the similarity by graph distance l_{gd}. Li use a factor α to tune the graph distance. The transfer function is defined as:

$$sim_1(ws_i, ws_j) = g_1(l_{gd}) = e^{-\alpha \cdot l_{gd}} \tag{8}$$

In Figure 9, S1 shows when $\alpha = 0.20$, computed similarities achieve the highest correlation with human judgments on training set D_1. Using as the optimal parameter the similarities of word pairs in testing set D_0 are calculated and their correlation with the human judgments is found to be 0.7972.

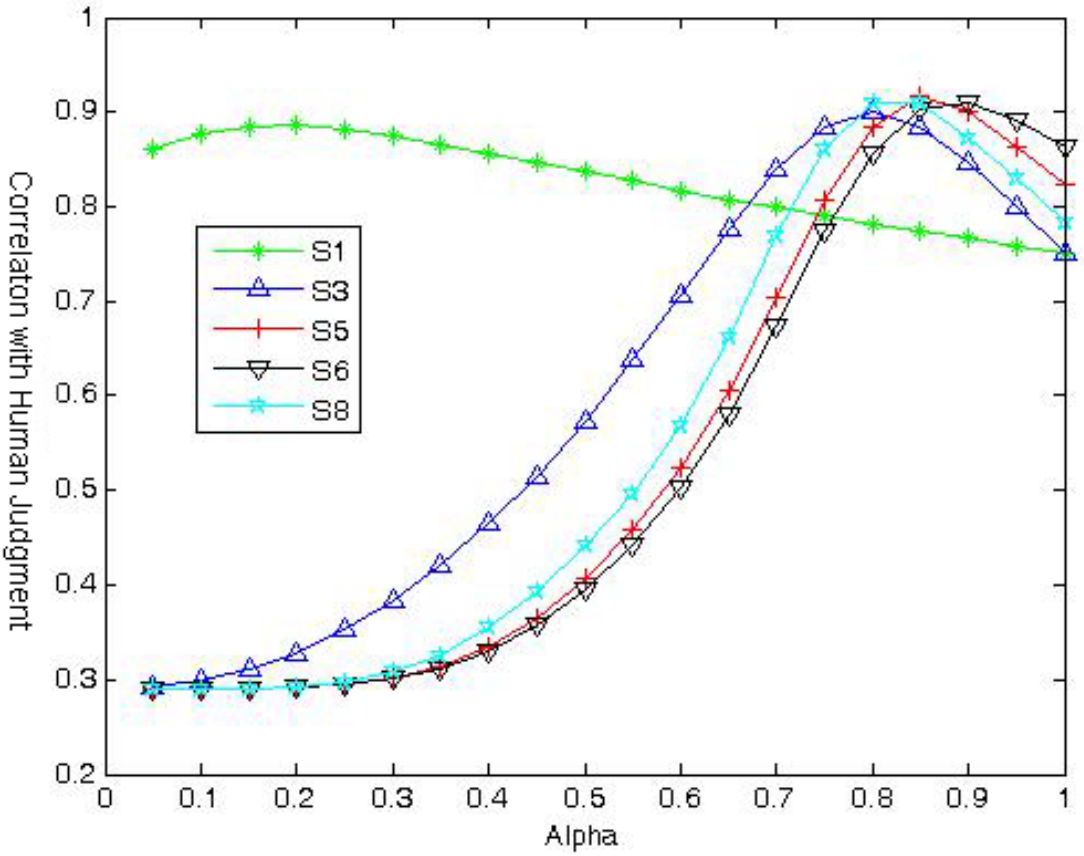


Figure 9: Correlation of one parameter strategies with MC human judgments

Strategy 2: We repeat Li's best strategy (Strategy 4) as our second strategy for comparison. This strategy considers both the shortest graph distance l_{gd} and SpecLev of their LCA $slev_{lca}$. It introduces a monotonically increasing function with respect to the Specification Level:

$$g_2(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Thus, the similarity function is defined as:

$$sim_2(ws_i, ws_j) = g_1(l_{gd}) \cdot g_2(slev_{lca})$$
$$= e^{-\alpha \cdot l_{gd}} \cdot \frac{e^{\beta \cdot slev_{lca}} - e^{-\beta \cdot slev_{lca}}}{e^{\beta \cdot slev_{lca}} + e^{-\beta \cdot slev_{lca}}} \quad (9)$$

This strategy has two tuning factors α and β. Factor α is used to model the impact of the graph distance to the similarity of words, and factor β is used to model the influence of the SpecLev of LCA. As shown in Figure 10, when $\alpha = 0.20, \beta = 0.3$, the computed similarities attain the highest correlation with the human judgments on training set D_1. Using these optimal parameters, we calculated the semantic similarities of word pairs in D_0 and found the correlation with the human judgments to be 0.8078.

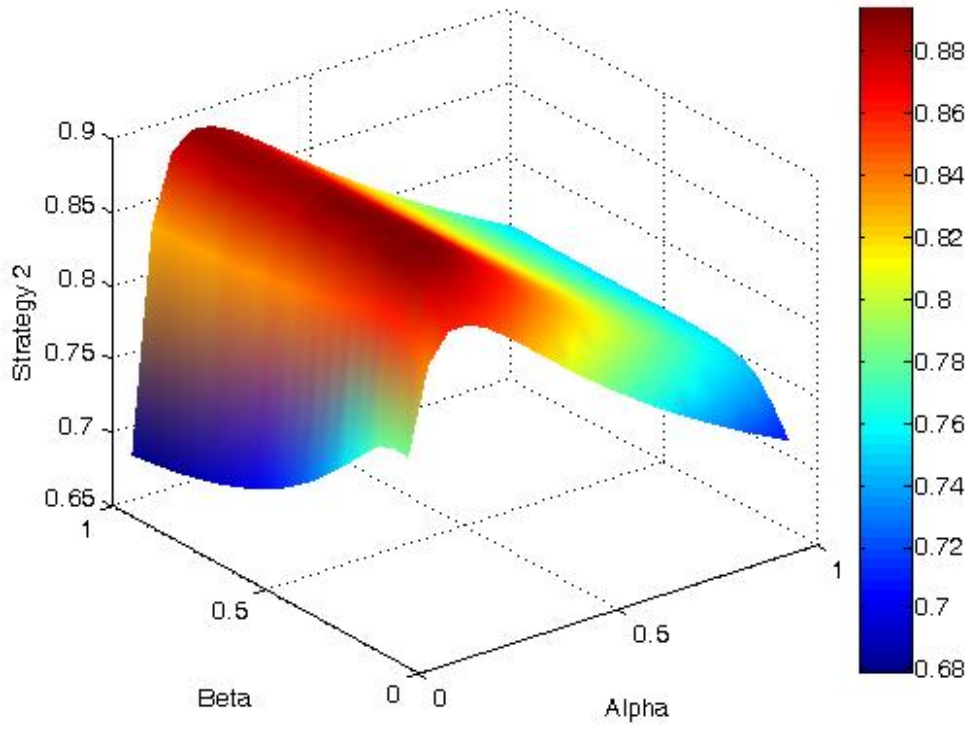


Figure 10: Correlation of Li's Best Method Strategy2

Strategy 3: This strategy is our first trial of weighted edge distance. We still use the monotonic increasing function g_1 from Strategy 1, but we replace the graph distance l_{gd} with our weighted edge distance l_w. It is worth noting that Li uses one specific factor α to tune the graph distance l_{gd}, but our method doesn't need the tuning factor because the value of weighted decreasing rate α is used for the tuning task.

$$sim_3(ws_i, ws_j) = g_1(l_w) = e^{-l_w} \tag{10}$$

Differing from Strategy 1, our weighted edge approach naturally adopts the non-linear mechanism, without needing an additional parameter to adjust the graph distance. As shown in Figure 9 S3, computed similarities have the highest correlation with the human judgments on training set D_1 when $\alpha = 0.80$. Using this parameter, the similarities for word pairs in testing set D_0 are calculated and their correlation with the human judgments is 0.8181. Clearly the result is better than both Strategy 1&2, especially Strategy 2 is Li's best strategy. This experimental study shows that our weighted edge approach model the human perception better than existing methods.

Strategy 4: To further compare with Li's Strategy 2, we replace the graph distance l_{gd} with our weighted edge distance l_w. The similarity function is as follows:

$$\begin{aligned} sim_4(ws_i, ws_j) &= g_1(l_w) \cdot g_2(slev_{lca}) \\ &= e^{-l_w} \cdot \frac{e^{\beta \cdot slev_{lca}} - e^{-\beta \cdot slev_{lca}}}{e^{\beta \cdot slev_{lca}} + e^{-\beta \cdot slev_{lca}}} \end{aligned} \tag{11}$$

When $\alpha = 0.8, \beta = 1$, this strategy yields the highest correlation between the computed similarities and the human judgments on training set D_1 shown in Figure 11.

Using these parameters to calculate the similarities of word pairs in testing set D_0, their correlation with the human judgments is found to be 0.8182.

Both Strategy 3 and Strategy 4 use the same α value. However, adding an extra parameter in Strategy 4 does not show much performance gain in terms of the correlation with human judgments. It confirms that with the weighted edge approach, it is unnecessary to use two parameters to calculate the semantic similarity.

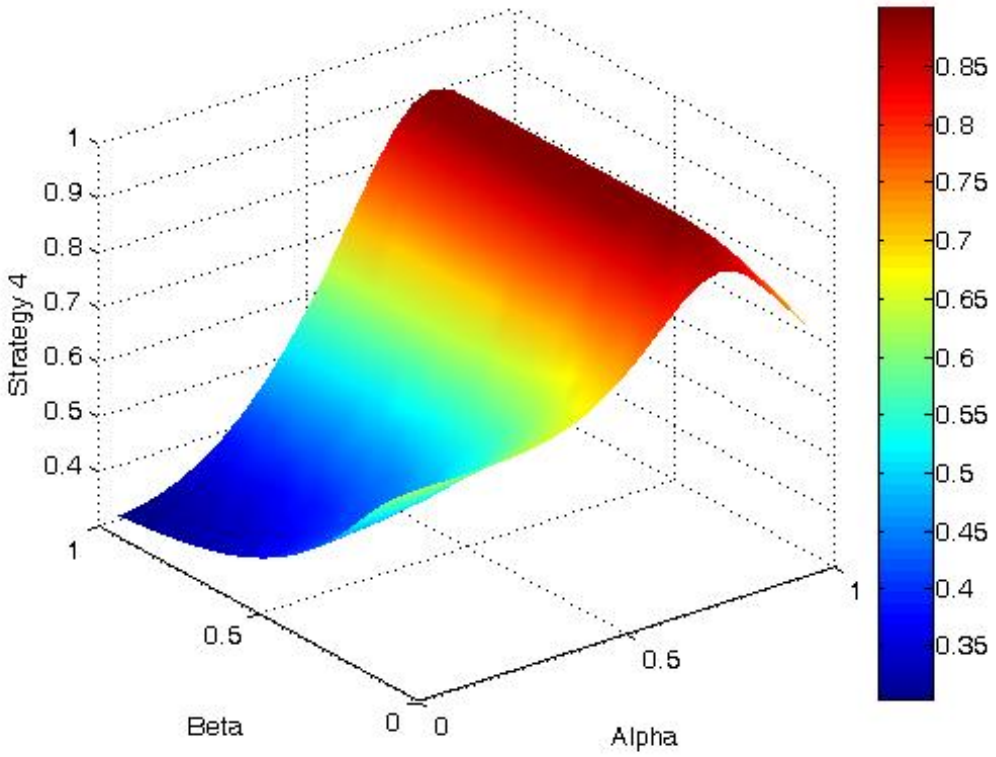


Figure 11: Correlation of Strategy 4

Strategy 5: This strategy uses hyperbolic secant (Sech) function to calculate the similarity:

$$g_5(x) = \sec h(x) = \frac{2}{e^x + e^{-x}}$$

$$sim_5(ws_i, ws_j) = g_5(l_w) = \frac{2}{e^{l_w} + e^{-l_w}} \tag{12}$$

As illustrated in Figure 9 S5, when $\alpha = 0.85$, the computed similarities have the highest correlation with human judgments on training set D_1. Using this parameter to calculate the similarities of word-pairs in testing set D_0 we found their correlation with the human judgments to be 0.8111. Again, this strategy is better than Li's strategy.

Strategy 6: In this strategy, we use the Hyperbolic Tangent Cardinal (*Tanhc*) function as our non-linear transfer function:

$$g_6(x) = \tanh c(x) = \begin{cases} \dfrac{e^x - e^{-x}}{(e^x + e^{-x}) \cdot x}, & x \neq 0 \\ 1, & x = 0 \end{cases}$$

$$sim_6(ws_i, ws_j) = g_6(l_w) = \begin{cases} \dfrac{e^{l_w} - e^{-l_w}}{(e^{l_w} + e^{-l_w}) \cdot l_w}, & l_w \neq 0 \\ 1, & l_w = 0 \end{cases} \tag{13}$$

As shown in Figure 9 S6, when $\alpha = 0.9$, the computed similarities have the highest correlation with human judgments on training set D_1. Using this parameter to calculate the semantic similarities of word pairs in testing set D_0, we found their

correlation with human judgments to be 0.8247, highest achieved so far. This confirms that a better non-linear function can improve the semantic similarity measure.

Strategy 7: This strategy is used to test whether combining the effects of two transfer functions can improve the performance. Here, the semantic similarity is measured by the linear combination of Strategy 5 and Strategy 6. That is,

$$sim_7(ws_i, ws_j) = \beta \cdot g_5(l_w) + (1-\beta) \cdot g_6(l_w) \tag{14}$$

An additional parameter β is used to weigh the values obtained by Strategy 5 and Strategy 6 respectively.

As shown in Figure 12, when $\alpha = 0.85, \beta = 1$, the computed similarities have the highest correlation with human judgments on training set D_1. Using these parameters to calculate the similarities of word pairs in testing set D_0, their correlation with human judgments is found to be 0.8111. It means a linear combination of Strategy 5 and Strategy 6 cannot improve the performance of semantic similarity measure. Actually, since the optimal β value is 1, this strategy is essentially the same as Strategy 5.

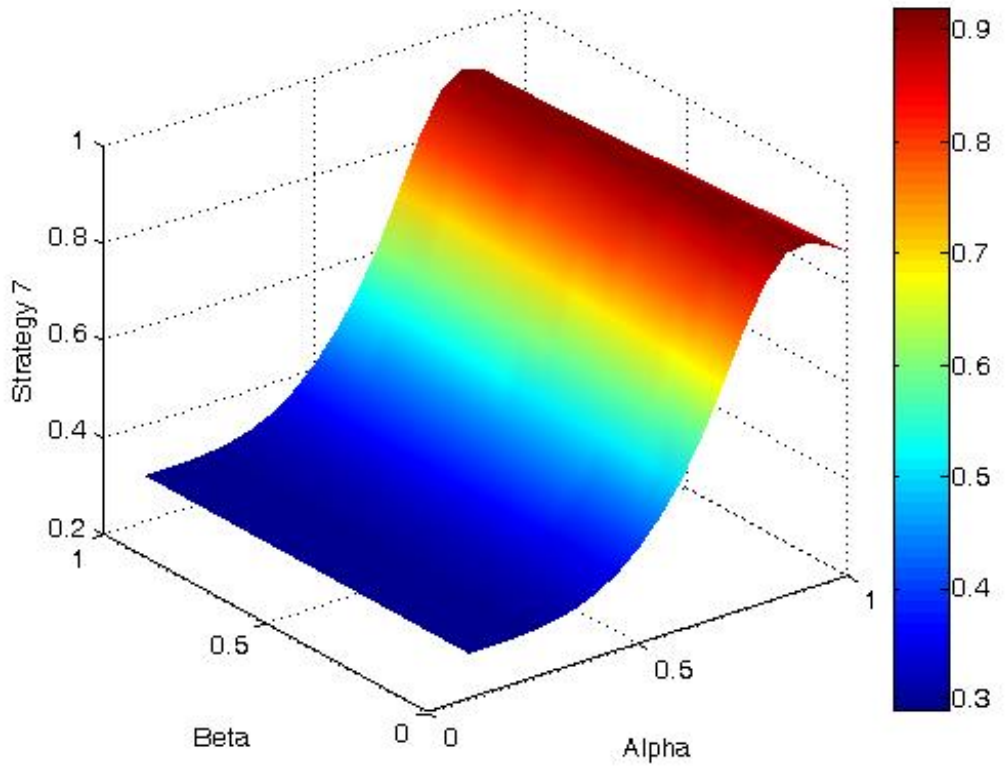


Figure 12: Linear combination of sech and tanhc in Strategy 7

Strategy 8: The final strategy combines Strategy 5 and Strategy 6 by multiplying the two hyperbolic functions. The similarity is calculated as:

$$sim_8(ws_i, ws_j) = g_5(l_w) \cdot g_6(l_w) \tag{15}$$

As depicted in Figure 9 S8, when $\alpha = 0.85$, the computed similarities have the highest correlation with the human judgments on training set D_1. Using $\alpha = 0.85$, we calculate the similarities of word pairs in testing set D_0 and found their correlation with human judgments is 0.83503, which is the highest correlation among all strategies tested.

3.5.3. Experimental Discussion

The correlations between the computed similarity values and the human judgments on testing set D_0 using four different strategies are summarized in Table 5. All our strategies (3-8) outperformed Li's best strategy (1, 2). Especially, the Strategy 8, a combination of *Sech* and *Tanch* transfer functions, achieves the best result.

Table 5: Correlations between WEST similarity and human judgments on testing set

Strategy	S1	S2	S3	S4	S5	S6	S7	S8
Correlation.	0.797	0.808	0.818	0.818	0.811	0.825	0.811	0.835

To better study the result of our approach, we record our semantic similarity data of all eight strategies and compare them with Miller-Charles human judgments in Table 6 and Table 7. Our experiments confirm that the distance-based methods are effective and accurate in measuring the semantic similarity of words when considering three factors: the graph distance of the words, the SpecLev of their LCA, and the SpecLev difference of these words. Our weighted edge model seamlessly integrates the three factors together and the similarity value can be easily tuned by a single parameter when adapting transfer function, instead of two parameters used by Li's best strategy.

Table 6: Comparison of S1-S4 on D0 testing dataset with MC Human Rating

Word1	Word2	MC Rating	Sim 1 a=0.2	Sim 2 a=0.2 b=0.3	Sim 3 a=0.8	Sim 4 a=0.8 b=1
cord	smile	0.13	0.14	0.04	0.01	0
rooster	voyage	0.08	0.01	0	0	0
noon	string	0.08	0.11	0.03	0.01	0
glass	magician	0.11	0.17	0.12	0.04	0.04
monk	**slave**	**0.55**	**0.45**	**0.43**	**0.41**	**0.41**
coast	forest	0.42	0.37	0.2	0.07	0.06
monk	oracle	1.1	0.25	0.23	0.24	0.24
lad	**wizard**	**0.42**	**0.45**	**0.43**	**0.39**	**0.39**
forest	graveyard	0.84	0.2	0.11	0.03	0.03
food	rooster	0.89	0.05	0.02	0	0
coast	hill	0.87	0.45	0.32	0.16	0.16
car	journey	1.16	0.03	0	0	0
crane	implement	1.68	0.45	0.41	0.32	0.32
brother	lad	1.66	0.45	0.43	0.39	0.39
bird	crane	2.97	0.55	0.54	0.72	0.72
bird	cock	3.05	0.82	0.81	0.87	0.87
food	fruit	3.08	0.17	0.09	0.02	0.02
brother	monk	2.82	0.82	0.81	0.87	0.87
asylum	madhouse	3.61	0.82	0.81	0.87	0.87
furnace	stove	3.11	0.17	0.14	0.08	0.08
magician	wizard	3.5	1	0.98	1	1
journey	**voyage**	**3.84**	**0.82**	**0.81**	**0.87**	**0.87**
coast	shore	3.7	0.82	0.68	0.66	0.66
implement	tool	2.95	0.82	0.78	0.77	0.77
boy	lad	3.76	0.82	0.81	0.85	0.85
automobile	car	3.92	1	1	1	1
midday	noon	3.42	1	0.99	1	1
gem	jewel	3.84	1	0.98	1	1

Table 7: Comparison of S5-S8 on D0 testing dataset with MC Human Rating

Word1	Word2	MC Rating	Sim 5 a=0.85	Sim 6 a=0.9	Sim 7 a=0.85 b=1	Sim 8 a=0.85
cord	smile	0.13	0	0.14	0	0
rooster	voyage	0.08	0	0.07	0	0
noon	string	0.08	0	0.13	0	0
glass	magician	0.11	0.03	0.18	0.03	0.01
monk	**slave**	**0.55**	**0.49**	**0.49**	**0.49**	**0.32**
coast	forest	0.42	0.08	0.27	0.08	0.03
monk	oracle	1.1	0.23	0.31	0.23	0.1
lad	**wizard**	**0.42**	**0.47**	**0.48**	**0.47**	**0.3**
forest	graveyard	0.84	0.03	0.19	0.03	0.01
food	rooster	0.89	0	0.11	0	0
coast	hill	0.87	0.2	0.36	0.2	0.09
car	journey	1.16	0	0.08	0	0
crane	implement	1.68	0.39	0.45	0.39	0.23
brother	lad	1.66	0.47	0.48	0.47	0.3
bird	crane	2.97	0.85	0.75	0.85	0.76
bird	cock	3.05	0.97	0.95	0.97	0.96
food	fruit	3.08	0.02	0.17	0.02	0
brother	monk	2.82	0.97	0.95	0.97	0.96
asylum	madhouse	3.61	0.97	0.95	0.97	0.96
furnace	stove	3.11	0.06	0.2	0.06	0.02
magician	wizard	3.5	1	1	1	1
journey	**voyage**	**3.84**	**0.97**	**0.95**	**0.97**	**0.96**
coast	shore	3.7	0.88	0.88	0.88	0.81
implement	tool	2.95	0.93	0.92	0.93	0.89
boy	lad	3.76	0.96	0.94	0.96	0.94
automobile	car	3.92	1	1	1	1
midday	noon	3.42	1	1	1	1
gem	jewel	3.84	1	1	1	1

3.5.4. Comparison with Li's Method

Our method performs better than Li's best strategy due to two reasons: (1) Our weighted edge distance model embedded the concept of *SLD* counting in the difference between inheriting and categorizing relationships, which is coherent with human perception. For example, word pairs in the testing set D_0 as shown in Table 6 and Table 7, "monk-slave" and "lad-wizard", have the same graph distance of 4. The SpecLevs of both pairs' LCA are 6. However, the *SLD* between "monk-slave" pair is greater than that of "lad-wizard". Thus, Weighted Edge Distance of "monk-slave" is less than that of "lad-wizard". By Strategy 7, the similarity value for "monk-slave" is 0.316 and "lad-wizard" is 0.296. These two results are consistent with MC human judgments, where the ratings for "monk-slave" and "lad-wizard" are 0.55 and 0.42 respectively. However, Li's strategy cannot distinguish the SLD, calculating the same similarity value for both word pairs.

(2) The second reason can be contributed to new hyperbolic transfer function, which matches with human perception more accurately in transferring weighted edge distance into similarity value. For example, in testing dataset, the MC human judgments of "monk-slave" and "journey-voyage" are 0.55 and 3.84 respectively, scaling from 0 (least similar) to 4 (exactly the same). The similarity values computed by our Strategy 7 are 0.316 and 0.957 respectively, compared with Li's strategy's 0.425 to 0.811. Obviously, our results are more consistent with human judgments than that of Li's.

3.5.5. The Impact of WordNet Evolution

As demonstrated by our experimental studies, Li's best strategy has a correlation of 0.8078 with human judgments. However, in Li's original paper, the correlation was reported as high as 0.8914, a huge difference from our experimental studies. Based on our observation and some previous studies, we can safely state that the evolution of WordNet is the main cause of this difference.

Due to the evolution of WordNet, the graph distance and the SpecLev of the LCA acquired from WordNet vary from version to version. For example, the graph distance between 'rooster-voyage' is 23 in WordNet 3.0, which is used in our experiments, but Li's paper obtained a graph distance of 30 using WordNet 1.6. Similarly, the graph distance between 'furnace-stove' is 9 in our experiments and only 2 in Li's paper due to the difference of WordNet versions. Based on these observations, we are not surprised that the correlation between the computed similarity values and the human judgments obtained in our study is quite different from that claimed in Li's paper.

The study by Varelas et al. [58] further confirms our observation. Varelas repeated Li's experiments and found that the highest correlation between the computed similarities and the human judgments is only 0.82, much less than 0.8914 which was claimed in Li's paper and much closer to our experimental results. Although Varelas did not mention which version of WordNet they used, we guess they used WordNet 2.0 or 2.1, considering the fact that WordNet 3.0 was released in Dec, 2006 and WordNet 2.1 Windows version was released in Mar, 2005.

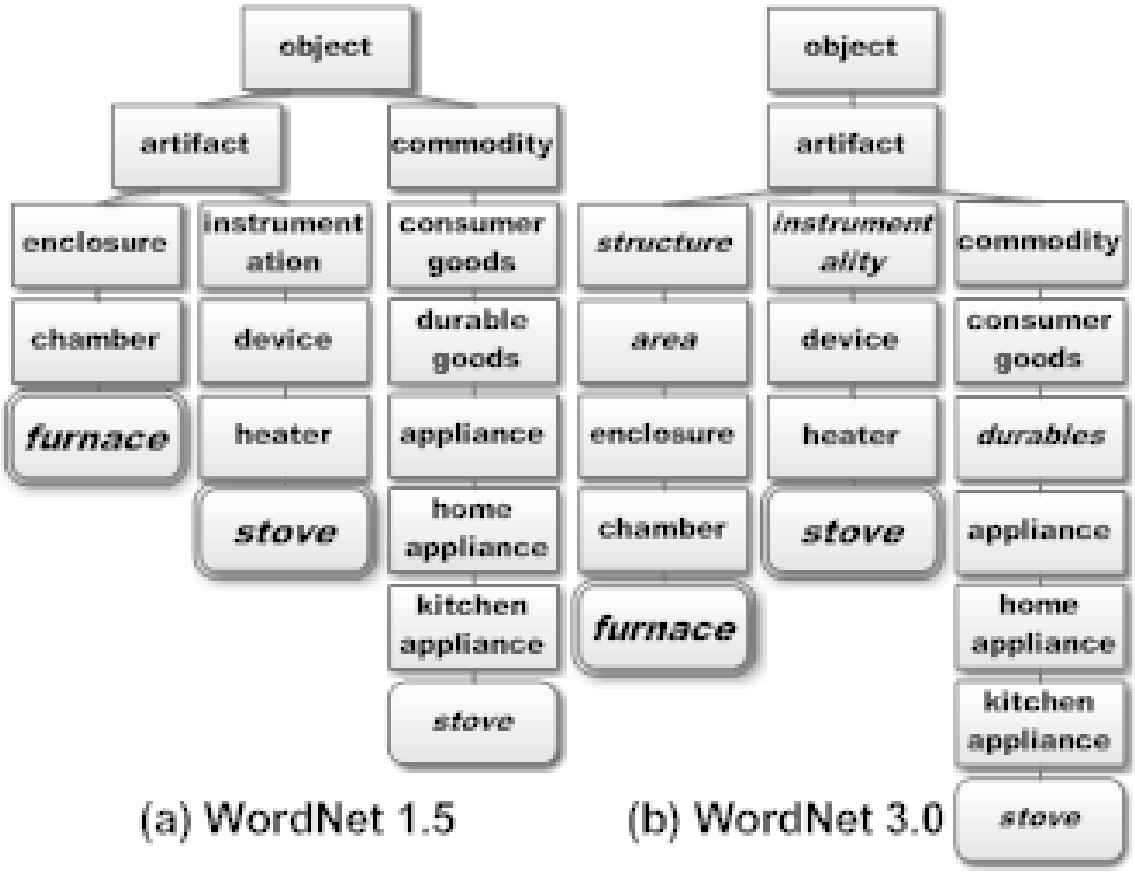


Figure 13: The evolution of WordNet structure

The structure change of WordNet has also been reported in some earlier studies. The WordNet structure in Figure 13(a) was illustrated by Jiang et al [50] who used WordNet 1.5 in their research. They discovered that the pair ‘furnace-stove’ was given high similarity values in human rating, whereas a very low rating (second to the lowest) was found using their method. They checked the WordNet hierarchy and found the shortest path of “furnace-stove” has a length 7. In Li’s paper which used WordNet 1.6, the same word pair has a very short graph distance 2. In WordNet 3.0, the shortest graph distance between this pair increases to 9 as shown in Figure 13(b).

Table 8 lists the correlations between the computed similarities and human judgments using Li's method under different WordNet versions.

Table 8: Li's method under WordNet versions

	Li, 2003	**Varelas, 2005**	**Us, 2009**
WordNet Version	1.6	2.0/2.1	3.0
Correlation	0.8914	0.82	0.8078

3.5.6. Comparison with IC-based approaches

Jiang [50] claimed his highest correlation is 0.8282 in his paper. However, that result was tuned to adapt the specific MC dataset. If they used the experimental methods as Li's and ours, their experimental result would be much more reliable and trustful.

Li tried to further improve the correlation between the computed similarity values and human judgments by combining the information content with graph distance in similarity measures, but found that the performance was degraded. Therefore, it is reasonable to believe that combining the information content and graph distance in measuring the semantic similarity of words may not improve the performance. Repeating Varelas' work, we also applied the WordNet similarity module implemented by Ted Pedersen [59] to calculate the correlations between the human judgments and the computed similarity values obtained by methods proposed by Resnik [49], Jiang [50] and Lin [51]. The only difference is that we use WordNet 3.0.

As shown in Table 9, Jiang's method and Lin's method, which used difference strategies to combine information content with graph distance, performed worse than Resnik's method, a pure information content-based method. Although an implementation

issue, explained Ted Pedersen himself, that “zero information content values in the denominator are handled in a special way in case of the Jiang-Conrath and the Lin measures”. The Perl module implementation returns extremely large number when two words are in the same synset, such as ‘hill-mound’ returns 12876699. The miscalculation greatly degraded the expected correlation performance. Another reason for Jiang’s poor performance is due to the Perl WordNet similarity module’s implementation chooses a simplified form ignoring the depth and density factors which further corrupt the expected correlation accuracy.

Table 9: Comparison with IC-based approaches

Method	**Type**	**G.Varelas, 2005**	**Us, 2009**
Resnik	Information Content (IC)	0.79	0.8124
Lin	Normalized IC	0.82	0.7517
Jiang	Hybrid	0.83	0.6900
Li	Graph Distance & IC	0.82	0.8078
Our method	Weighted Edge	-	0.8350

3.5.7. Computational Cost Analysis

Retrieving the Least Common Ancestor in the WordNet is the Least Common Ancestor (LCA) problem, which is the same as Range Query Minimum (RQM) problem. Harel and Tarjan [60] showed an algorithm to find two nodes’ LCA in constant time with a linear preprocessing of the tree structure. Bender and Farach-Colton [61] presented a simplified algorithm with O(n) preprocessing time and constant time to obtain LCA under a tree structure. Bender et al [62] proposed an algorithm solving LCA problem on Direct Acyclic Graph (DAG) with $O(n^{2.688})$ preprocessing time and $O(1)$ query time. Since

the WorldNet hierarchy is constructed as DAG rather than tree, we can achieve this $O(n^{2.688})$ pre-processing time and $O(1)$ execution time to retrieve the LCA of two synsets in WordNet graph.

Steyvers [63] illustrates that Zipf's (Power Law) Distribution applies not only to word frequency, but also to the number of senses of English word. That is, most words have a small amount of senses, and only a few words have a large amount of senses. Empirically, those words with many senses are coherent with those high frequency words which would be trimmed if using stop-lists. Though we need to iterate $O(n^2)$ time to find the best sense-pair when measuring a word-pair, the expectation for the number of senses of each word is low – thus - it is still applicable in real-application.

3.6. Weighted Edge Similarity Web Tools

3.6.1. Web Architecture

Providing Web services or application packages for word similarity measures will benefit researchers in related research fields. Existing web tools or packages for word similarity measures are limited. MSRs [64] is an implementation of word similarity methods based on several large text corpora. A Web Server is publicly available at http://cwl-projects.cogsci.rpi.edu/msr, which measures word semantic relatedness based on corpus such as Google, Wikipedia, New York Times, and so on. WordNet::Similarity [59] is a powerful Perl Module developed by Ted Pedersen et.al. They have pre-computed information content from the *British National Corpus* (World Edition), the Penn

Treebank (version 2), Brown Corpus, the complete works of Shakespeare, and SemCor (with and without sense tags). A web Interface is provided embedding eleven different word similarity methods under both Graph-distance and IC categories. UMLS:: Similarity [65] is a recent proposed Perl Module calculating the semantic similarity between concepts in Unified Medical Language System (UMLS) using several previously developed similarity measures such as Wu [46], Leacock [66], and Nguyen [67].

To disseminate our proposed new method for word similarity measure, we have built and published a set of web-based tools and services online. We call our tool set WEST--Weighted-Edge based Similarity Measurement Tools [68]. This section presents the design and analysis of this WEST environment. We will introduce the architecture of the system and the implementation details of weighted edge approach.

The WEST environment is built upon Client-Server architecture. The web server is deployed at the School of Computing, Clemson University, South Carolina. The details of WEST architecture are shown in Figure 14.

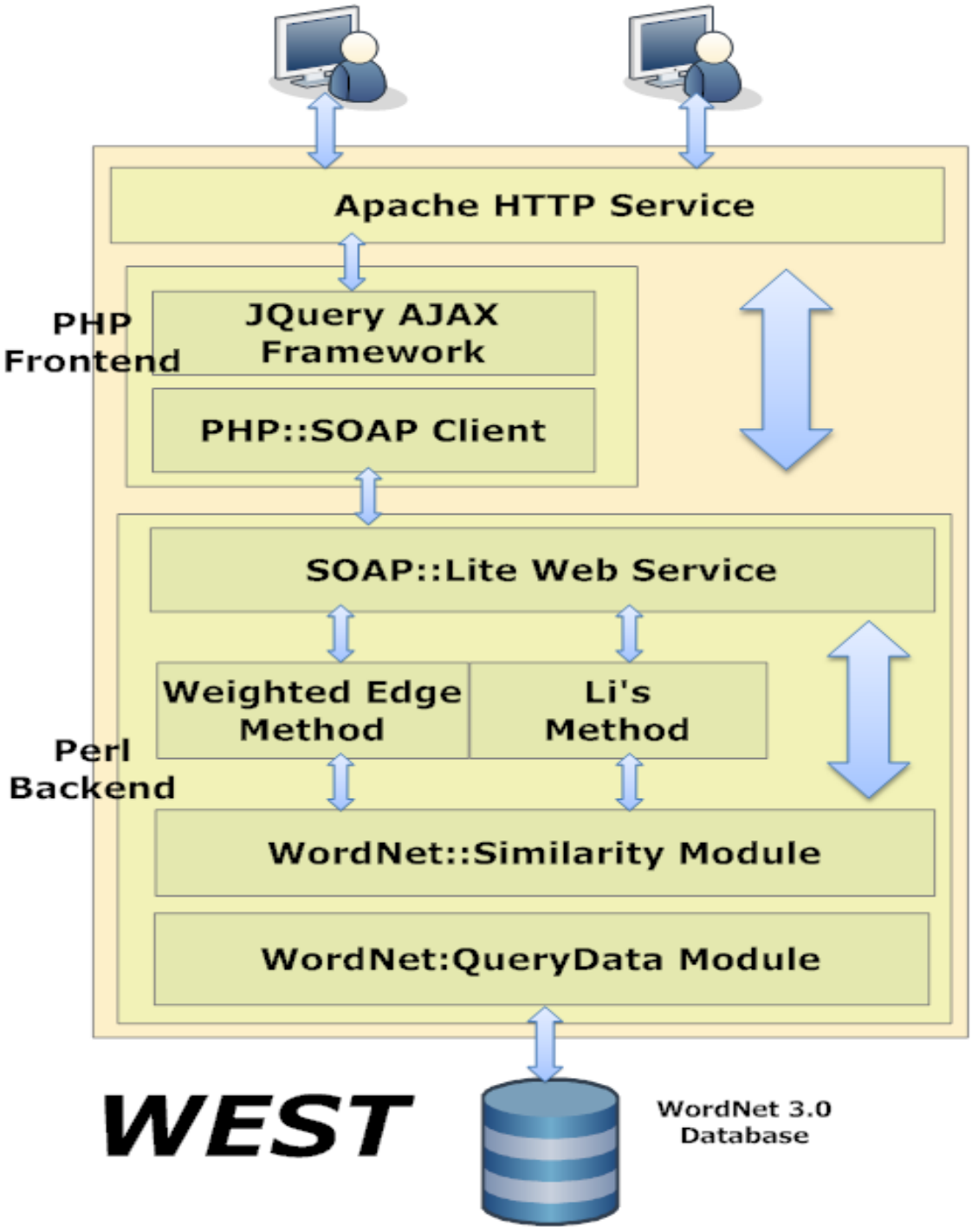


Figure 14: WEST Architecture

The operating system of the server is CentOS; Apache server provides the HTTP service for the whole environment. The backend computational measurement is written in Perl. WordNet::QueryData and WordNet::Similarity modules are used in our project to access the WordNet database, and provide existing similarity measurements from previous studies. SOAP::Lite module is employed providing the SOAP web service to both frontend and the public. The frontend user-interface is coded by PHP. PHP::SOAP client exchanges data from the Perl SOAP server. JQuery, the most popular Ajax Framework, is employed to provide an interactive experience between user and the web-environment.

3.6.2. Implementation

A detail implementation of Weighted Edge approach is introduced in this section. There are seven steps to conduct the similarity measurement for any word pair (w_i, w_j).

(1) Stemming: We need to pre-process any given word before conducting similarity measurement. A simple WordNet morphology function wrapped by Similarity Module is used to stem word into original form e.g. “dogs”->”dog”.

(2) Part-of-Speech (PoS): In WEST, we need to test the PoS of any given word to ensure the word is a noun. The WordNet hypernym relationship only applies to noun and verb, not to adjective, adverb or the others. Since nouns are widely acknowledged play the most decisive role in information retrieval applications, currently WEST is focusing on noun to prove its effectiveness. However, Weighted Edge method works the same for verbs.

(3) LCA Selection: For each sense-pair of a word-pair, we use PathFinder from WordNet::Similarity module to retrieve its LCA. Among all retrieved LCA, we keep a record on the highest SpecLev and corresponding sense-pair. If there are two or more such LCAs, we take the LCA on the shortest graph distance path. Thus, we retrieve the sense-pair of a word-pair with highest similarity and its LCA sense.

(4) SpecLev Retrieval: Level function from WordNet:: QueryData module is used to retrieve the SpecLevs of the three target senses.

(5) Weighted Edge Distance: Equation 15 is used to calculate the Weighted Edge Distance. We optimize the calculation by pre-calculating the Weighted Edge Distance from all SpecLev to its root (SpecLev 0) and store them into a 2-dimentional array for every Weighted Decreasing Rate. Thus, the computation only spends constant time.

(6) Transfer Function: After the Weighted Edge Distance has been calculated, we can apply the transfer functions in section 3.5.2 to change the Weighted Edge Distance into similarity value.

(7) Web Service: SOAP::Lite Module is used to wrap the Weighted Edge interface into SOAP web service.

3.7. Summary

This chapter presents a novel WordNet-based method to measure semantic similarity of a word pair and provide a set of Web-based tools and APIs that can be used by public. Weighted Edge approach is based on an important observation that humans are more sensitive to the semantic difference caused by the categorization than by specification. Therefore, people view word pair separated by specification more similar than those separated by categorization. Our weighted edge distance model merges the specification level difference of a word pair and the specification level of its least common ancestor together. Based on this new model and a set of improved non-linear transfer functions, our method's result reaches the highest correlation against Miller-Charles's human similarity judgment by far.

Chapter 4

Ontology Graph based Query Expansion

4.1. Motivation

Since the beginning of the new millennium, the explosively growing biomedical data has made it difficult for the researcher to keep up-to-date with ongoing research. It is important to capture the latest biological discovery from literature which demands for an efficient and effective *biomedical information retrieval* (BIR) system. Though many existing information retrieval techniques can be directly used in BIR, BIR differs from traditional information retrieval in its widely used biomedical terms and abbreviations which are not presented in traditional thesaurus. One of the difficulties in BIR is to increase the recall and precision performance in searching *MEDLINE* database. MEDLINE is a large bibliographic database that contains more than 18.9 million documents (by July 2011) of medical journals and articles. NCBI's *PubMed* system is the most widely used web system for searching MEDLINE.

However, effectively querying MEDLINE by PubMed is not an easy task for normal users. It is widely reported [6, 7] that normal users do not utilize the system as

effectively as experts. Those inexperienced searchers either fail to employ the best query terms or fail to effectively apply Boolean expressions in the query statement [8].

4.1.1. Related Works

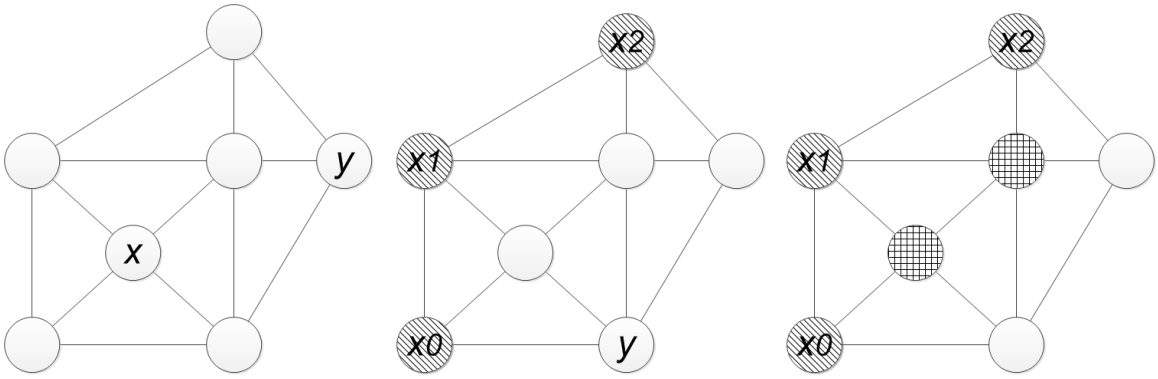
Using MEDLINE to perform biomedical information retrieval has been studied since early 1990s [9-11]. Those early studies observed that using controlled vocabularies such as MeSH offer no advantages in retrieval performance over free-text. The poor performance is caused by a number of potential reasons such as missing concepts and incomplete synonym sets [12]. Srinivasan [13, 14] observed that *pseudo relevance feedback* (PRF) based query expansion on MeSH vocabulary improved the retrieval performance. Yoo [15] and Abdou [16] re-designed/modify the terms weight scheme found by PRF. However, since PubMed doesn't sort matched documents by relevance, the PRF strategy might not apply properly into PubMed.

All above query expansion methods have a common weakness that they only used one controlled vocabulary - MeSH. The problem of the ineffective searching of MEDLINE is caused by its heavy usage of the MeSH vocabulary in its indexing and user-querying components. There are 26,142 descriptors, 83 qualifiers, over 177K assisting entry terms and over 199K supplementary concept records in MeSH 2011; but only descriptors and qualifiers are used in indexing MEDLINE. In comparison, NLM *Metathesaurus* 2010AB covers 2.3 million biomedical concepts. The primary disadvantage of the MEDLINE/PubMed system is that it indexes millions of documents with less than 1.1% of the available biomedical vocabulary.

Matos [69] invested in query expansion for gene related publication which expands genes to its related proteins, pathways and diseases, but it is not a general method. Taylor [17] expanded the query with inter-concept relationship by reformatting the query into a semantic graph. The problem of this method is over-emphasizing the concepts with inter-relations; besides it is computationally expensive in building the semantic graph for indexing documents.

Recently, Personalized PageRank based methods are applied in two natural language processing fields. In 2009, Agirre and Soroa [70] first proposed the application of Personalized PageRank in *Word Sense Disambiguation* (WSD) using WordNet as knowledge base. Later they made further study on biomedical WSD [71] using Metathesaurus as knowledge base. Personalized PageRank is also used to measure word/text semantic similarity. Agirre and Alfonseca [72] used Personalized PageRank to compute word similarity using WordNet as knowledge base. Later they compared their methods using various knowledge bases [73]. Ramage applied a similar random walk method to measure text semantic similarity [74].

4.1.2. Word Similarity and Query Expansion Problem



(a) (b) (c)

Figure 15: Relationship between word similarity and query expansion problem

Given a graph $G = (V, E)$ in Figure 15(a), let's assume an arbitrary word similarity function $Sim(x, y)$ can be used to calculate the similarity between node x and node y. We define *Accumulated Similarity* in Figure 15(b) between node set $X = \{x_0, x_1, ..., x_n\}$ and node y as Equation 16:

$$AS = \sum_{i=0}^{n} Sim(x_i, y) \tag{16}$$

With the above definition, the query expansion problem can be represented by word similarity problem in Figure 15(c): Given a graph $G = (V, E)$, an arbitrary similarity function $Sim(x, y)$ and a node set $X = \{x_0, x_1, ..., x_n\}$, query expansion aims to select top *K* nodes with the largest Accumulated Similarity from the rest of the nodes in the graph.

It is worth noting that for the query expansion problem, we need to use accumulated similarity rather than the maximum similarity $\underset{0 \le i < n}{Max}\{Sim(x_i, y)\}$ to prevent query drifting [42-44, 75]. Query drifting can cause the degradation of the search performance and is the worst case for query expansion.

4.1.3. Contributions

In this chapter, we propose and evaluate a novel and effective ontology graph based query expansion scheme for biomedical search engine by utilizing a subset of UMLS Metathesaurus. Our contributions are six-folds. First, this novel query expansion method is conceptually different from previous techniques as of our knowledge. Second, the query expansion analyzes the whole context of user query rather than individual terms in the query. Third, unlike many previous studies which utilize only MeSH, our method can employ multiple controlled vocabularies from Metathesaurus for indexing/searching. Fourth, we showed that generalized biomedical concepts may degrade retrieval performance. Fifth, we designed a systematic method to eliminate the mapped generalized biomedical concepts and populate closely related specialized concepts resulting in significant increase in the relevance of retrieval results. Sixth, we demonstrate that query expansion based on ontology graph is more stable than that based on pseudo relevance feedback because sorting the retrieved documents by relevance is found to be often inaccurate.

4.2. Personalized PageRank Algorithm

The *PageRank* algorithm, a method for computing the relative rank of web pages based on the linkage structure of the web, was introduced in [76, 77] and has been widely used since then. The fundamental motivation underlying the basic foundation of PageRank algorithm is recognition and use of the fact that important pages are almost always linked to many other important pages.

Consider a *random surfer* who begins at web page and executes a random walk on the web as follows: at each time step, the surfer proceeds from his current page to a randomly chosen web page that it hyperlinks to. As the surfer proceeds in this random walk from node to node, he visits some nodes more often than others; intuitively, more frequently visited nodes are those with many in-links coming from other frequently visited nodes. For a detailed review of PageRank computing, see [78-81].

Let $G = (V, E)$ be a directed graph with vertices $V = (v_1, ...v_N)$ where the nodes represent web pages and directed edges E represents the directed hyperlinks. Let n be the total number of pages, the edges E are given by a (often sparse) nonnegative matrix $M_{n \times n}$, where $M_{ij} = 1 \, iff$ there is a direct link from vertex v_i to vertex v_j and equal to 0 otherwise. Let $\deg(i)$ denote the out-degree of vertex v_i. For pages with non-zero number of out-links $\deg(i) > 0$, the rows of M can be normalized into a *row-stochastic* matrix by $P_{ij} = M_{ji} / \deg(i)$, where the sum of components in each row is one. If $\deg(i) = 0$, we set the entire row component to zero.

Given a vertex v_i, let $In(v_i)$ be the set of vertices pointing to it, PageRank of v_i is defined as:

$$P(v_i) = c \sum_{v_j \in In(v_i)} \frac{1}{\deg(j)} P(v_j) + (1-c) \frac{1}{N} \tag{17}$$

where $c \in (0,1]$ is the so-called *damping factor* which ensures the *irreducibility* and *aperiodicity* properties so that the iterative power method can converge to principal eigenvector as solution. Note that a web page user follows one of the local out-links with

probability c and teleports to another random page with probability $1-c$. In this paper, we simply choose a heuristic damping factor value 0.85.

The PageRank score reflects a "democratic" principle in the sense that the user has no preference for any particular pages. However, a random surfer may have a set of preferred pages where he is more likely to be teleported to in real world. The algorithm can be modified to reflect biased user preference (such as bookmark pages), called *Personalized PageRank* [82], by replacing the uniform teleportation probability vector with non-uniform one. For an overview of recent personalization methods, see [79, 83].

We rewrite Equation (17) in terms of normalized teleportation probability vector $\mathbf{v}$. The calculation of PageRank Vector $\mathbf{P}$ is equivalent to:

$$\mathbf{P} = cM\mathbf{P} + (1-c)\mathbf{v} \tag{18}$$

The teleportation probability vector $\mathbf{v}$ is non-uniformly distributed in Personalized PageRank; thus the random web page user has a higher (teleportation) probability to jump back to the original page. Thus, the *Personalized PageRank Vector* (PPV) $\mathbf{P}$ represents the importance of the entire vertices effectively biased by the initial non-uniform teleportation probability vector.

4.3. Fundamental Notion

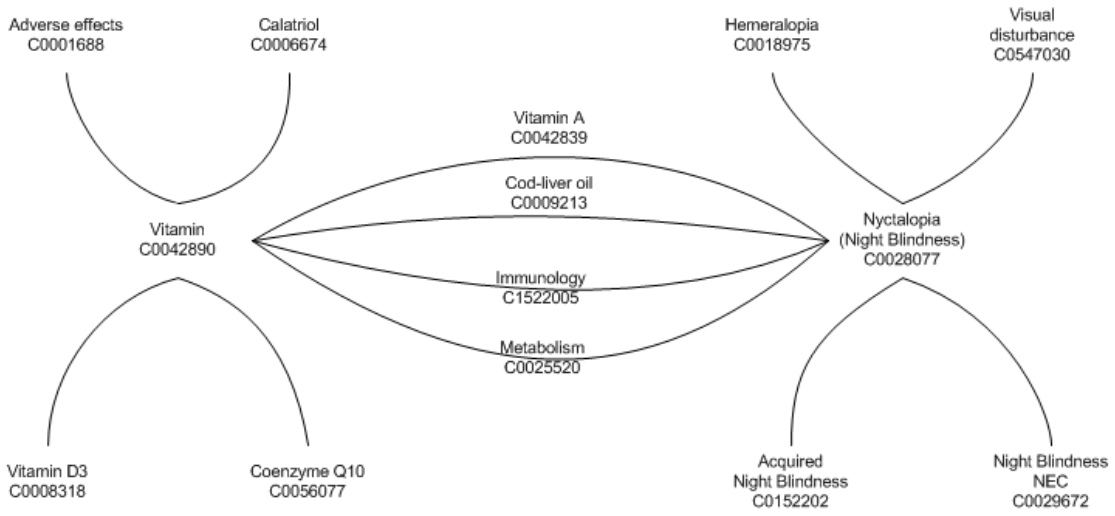


Figure 16: An example fraction of a biomedical ontology graph

Before we dive into the technical details, we want to explain the fundamental idea underlying the personalized PageRank algorithm in query expansion.

Let's assume a searching scenario in the first place. Given two concepts "Vitamin" and "Nyctalopia" as user input, a small portion of the sub-ontology graph is illustrated in Figure 16. For a better illustration of the graph, we choose one simple English word to represent each concept in the figure. Depending on the size of the ontology graph, there might be hundreds of concepts related to either "Vitamin" or "Nyctalopia", and tens of concepts related to both concepts. For query expansion, it is very straightforward to prefer those concepts linked to both "Vitamin" and "Nyctalopia" such as "Vitamin A". Although it is plausible to directly probe the two neighbor sets of each concept and compute its intersection, it is much complicated and computational

expansive if we consider the situation of either inter-relationships among multiple concepts or relationships separated by multiple hops.

By using personalized PageRank, we can imagine that the random surfer is teleported back to either “Vitamin” or “Nyctalopia” every time. Thus, “Vitamin” and “Nyctalopia” will have the highest probability distribution in the final Personalized PageRank Vector (PPV); followed by those concepts linked to (or near to) both “Vitamin” and “Nyctalopia” such as “Vitamin A”, “Cod-liver oil”. Those concepts linked to (or near to) only one concept are assigned the lower probability value. Of course, concepts far from both “Vitamin” and “Nyctalopia” are assigned the lowest probability value. The merit for personalized PageRank is that it naturally assigns higher value to those concepts linked to or near to more original concepts. Besides, it can treat concepts separated from original concepts by multiple hops with different value. In a word, by computing the PPV, we acquire the probability distribution of the concepts from the entire graph and the PPV serves as relational indicators for each concept to the original input concepts. Section 4.4.1 describes the construction of the ontology graph; section 4.4.2 introduces mapping input text to biomedical concepts; section 4.4.3 applies the personalized PageRank to compute PPV. Nevertheless, there is one problem if we directly use the rank in PPV into query expansion. Among all four concepts linked to both “Vitamin” and “Nyctalopia” in the Figure 15, concepts “Vitamin A” and “Cod-liver Oil” are certainly very interesting as expanded terms; however, “immunology” and “metabolism” are not. How can we evaluate “Vitamin A” and “Cod-liver Oil” higher than “immunology” and “metabolism”

in the query construction? We propose a weighted scheme in section 4.4.4 to solve this problem. Section 4.4.5 assembles the rest elements for building a search engine.

4.4. Ontology Graph based Query Expansion Method

The flow chart in Figure 17 shows the major steps of our method to construct a new expanded query. There are total five steps where each step is corresponding to a single subsection.

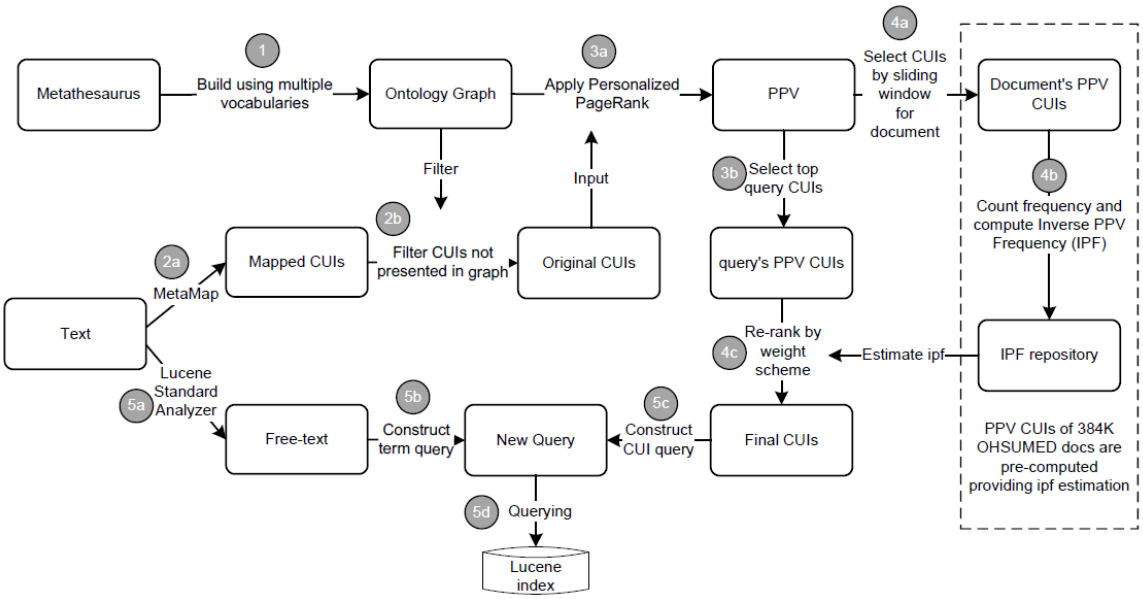


Figure 17: Flow chart describing the query expansion procedure

4.4.1. Ontology Graph Construction

The Metathesaurus of Unified Medical Language System (UMLS) [28, 29] is a large, multi-purpose, and multi-lingual vocabulary database containing information about biomedical related concepts, their various names, and their inter-relationships.

Each biomedical concept is identified by a distinctive id called *Concept Unique Identifier* (CUI), which is an eight character alpha-numeric string. We use CUI to represent each biomedical concept in this paper. Each CUI is associated with a set of lexical variants strings, called *concept name*. The concept name may refer to medical conditions, appendages, diseases, drugs, and others; it may be single term, phrase, or a string of terms. The `MRCONSO` table stores the entire CUIs and concept names.

The Metathesaurus includes many inter-concept relationships as well. Most of these relationships come from individual vocabularies. The others are either added by NLM during Metathesaurus construction or contributed by users to support certain types of applications. The inter-concept relationships are stored in the `MRREL` table. Many types of relationships are included such as parent/child, immediate siblings.

The construction of ontology graph matches Step 1 in Figure 17. An ontology graph is constructed using the information from `MRCONSO` and `MRREL` tables. The concepts are represented as vertices, and all the inter-concept relationships are represented as edges. The type of the inter-concept relationship is not distinguished so that there is no weight attached to the edges of ontology graph.

Table 10: Multiple vocabularies and #CUIs

Group	Acronym of Vocabulary	Full Name of Vocabulary	#CUIs
I	MSH	Medical Subject Headings	313,372
	SNOMEDCT	SNOMED Clinical Term	320,648
	CSP	CRISP Thesaurus, 2006	16,680
	AOD	Alcohol and other Drug	15,900
II	GO	Gene Ontology	54,453
	ICD10CM	Int'l Classification of Disease, 10^{th} edition, Clinical Modification	97,664

	NCI	NCI Thesaurus	81,455
	RXNORM	RxNorm Vocabulary	193,737
III	MTH	Metathesaurus MTH	138,003
	NCBI	NCBI Taxonomy	478,196
	RCD	Clinical Term Version 3	186,032

In our study, we used Metathesaurus 2010AB including total 2,381,619 concepts. Four major English vocabularies in Group I (MSH [30], SNOMEDCT [32], CSP, and AOD) with total 620,387 concepts are employed to build our *Origin* ontology graph. Eight vocabularies from Group I+II with total 988,490 concepts are used to construct *Medium* ontology graph. Finally, all eleven vocabularies from Group I+II+III with total 1,470,588 concepts are used to build the *Large* ontology graph.

Table 10 lists the full name and the number of concepts from each vocabulary. It is worth noting that we studied the difference of *Origin, Medium, Large* ontology graphs in chapter 4.5.5. The rest of the chapter only applies to the *Origin* ontology graph.

4.4.2. Mapping Text to CUI

The task of automatically mapping biomedical text to UMLS Metathesaurus is performed by *MetaMap* [84, 85], a supporting software tool provided by NLM. MetaMap uses a knowledge intensive approach based on symbolic, *natural language processing* (NLP) and computational linguistic techniques. MetaMap has been used in biomedical information retrieval and data mining applications, and automatic indexing of biomedical literature at NLM.

Shown in Step 2a in Figure 17, MetaMap first splits an input text into a set of noun phrases and generates the variants for each noun phrase where a variant essentially consists of one or more noun phrase words together with all of its spelling variants, abbreviations, acronyms, synonyms. Then, it maps a set of candidate CUIs containing one of the variants and computes a score for each candidate CUI by an evaluation function. Finally, it combines candidates involved with disjoint parts and re-computes the score based on the combined candidates. Those CUIs with highest score are selected as the best match to the input text.

Since only a subset of the Metathesaurus is used to build the ontology graph, we keep only those mapped CUIs that exist in the four selected vocabularies. Those CUIs are called *Original CUIs*, shown in Step 2b in Figure 17.

MetaMap2010 maps MEDLINE document's title, abstract, and query text to Metathesaurus CUIs. The Med-Post/SKR part-of-speech tagger and word sense disambiguation are enabled during the process.

4.4.3. Personalized PageRank on CUI

Recall the concept of Personalized PageRank from section 4.2. Given a part of biomedical related text, mapped CUIs produced by MetaMap can be used as the initial teleportation probability vector to compute *Personalized PageRank Vector* (PPV) defined in Equation 17 via power iteration.

Next, PPV is computed based on the Original CUIs as the teleportation probability vector on the ontology graph. We denote the top scored CUIs in the computed

PPV be *PPV CUIs*, noted as Step 3a in Figure 17. Scores of PPV CUIs are L1-normalized.

It's worth noting that Personalized PageRank ensures the Original CUIs are present and highly scored in the computed PPV CUIs.

PPV CUIs of query text are used as query expansion candidates. Since the query text is very short that only 2-4 Original CUIs are mapped for query in most of the case, we select a fixed top 500 scored PPV CUIs as candidates for each query, shown as Step 3b in Figure 17.

The PPV computation is performed by an open source C++ tool called UKB[1] [70], which is originally used to perform WSD.

4.4.4. Weight Scheme of PPV CUIs

The key value of our proposed ontology graph based method is to effectively and efficiently build the L1-normalized query PPV CUIs into expanded query.

However, there are two reasons why we cannot directly use the PPV CUIs into query expansion.

First, the scores are not very discriminative for direct usage in query expansion. The Personalized PageRank algorithm ensures the existence and high score of the Original CUIs ranked in the PPV CUIs. If we sort the PPV CUIs in descending order, the Original CUIs are distinguished from the rest PPV CUIs with high score and the score gaps between the two groups are large in most cases. The rest of the PPV CUIs have

much lower scores as well as tiny score gap between two consecutive CUIs. Thus, directly using PPV CUIs make trivial difference from simply using Original CUIs.

Second, the Personalized PageRank algorithm also guarantees that generalized concepts (more links) are scored higher than specialized concepts (less links). This phenomenon causes dozens of general medical concepts, such as 'disease' or 'therapy', frequently appeared and highly ranked in the PPV CUIs list.

To alleviate the problem, we propose a weighted scheme to compute a new weight w_i for each PPV CUI i in order to re-rank the PPV CUIs. Analogous to the classic *tf-idf* form in information retrieval, the query weight formula *ps-ipf* is defined as:

$$w_i = ps_i \cdot ipf_i \tag{19}$$

$$ps_i = s_i^{\alpha} \tag{20}$$

$$ipf_i = \max\{0, \log(\frac{N - n_i + 0.5}{n_i + 0.5})\} \tag{21}$$

The Equation (19) is a combination of two factors. The first factor ps_i is acronym for *PPV Score*, serving as term frequency: s_i is the L1-normalized PPV score of CUI i; and $\alpha \in [0,1]$ is a tuning parameter used to increase PPV score by decreasing α. The second factor is called *inverse PPV frequency* (IPF), which is analogous to inverse document frequency based on probabilistic ranking model [86], where N is the total number of computed PPVs in the collection, and n_i is the number of PPVs containing that specific PPV CUI i. In addition, plus .5 prevents the error when $N = n_i$.

[1] http://ixa2.si.ehu.es/ukb/

To statistically estimate IPF in Equation (21), we computed and indexed a large amount of PPVs from biomedical corpus to build an *IPF repository*. Shown as Step 4a in Figure 17, the PPV CUIs for document are computed using a sliding window method, different from the fixed top 500 query PPV CUIs for query text. Because of the title and abstract texts may have arbitrary length with various numbers of Original CUIs, a sliding window with size 100 is applied on the sorted PPV CUIs list to truncate the sequence when the difference in scores between the first and last CUI in the window drops below 5% of the highest-scoring PPV CUI.

In our study, we compute PPV CUIs generated from 348K OHSUMED documents to build the IPF repository shown as Step 4b in Figure 17. Thus, we can estimate the IPF by counting PPV frequency n_i for every CUI using Equation (21), shown as Step 4c in Figure 17.

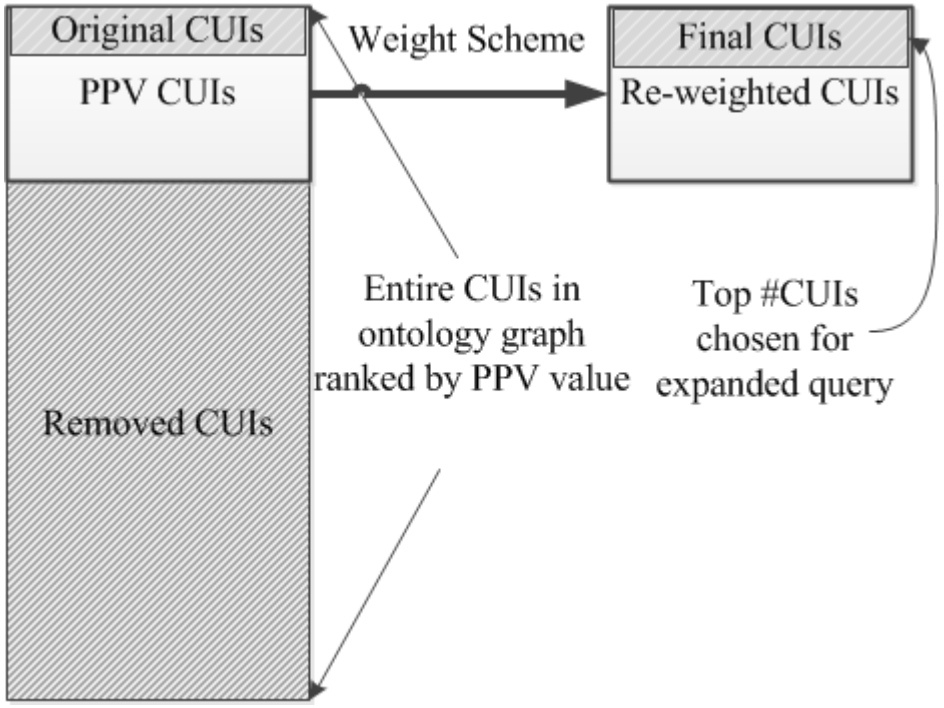


Figure 18: Weight scheme re-rank the order of CUIs

After the weights of all PPV CUIs are computed using Equation 19, we sort the query PPV CUIs again by selecting the top ranked k candidates, called *Final CUIs* in Figure 18 The computed weights of Final CUIs are divided by the highest weight for normalization so that those final weights are in the range $[0,1]$.

Finally, a boosting value b is used as an influence factor by multiplying the score of Final CUIs during the final query construction.

4.4.5. Document Indexing and Retrieval

To perform biomedical information retrieval efficiently, we use the popular Apache Lucene[2] Java search library version 2.9.4 to create local index for MEDLINE documents.

In the indexing stage, a modified Lucene standard analyzer with an enhanced stop-list[3] and Porter stemmer is used to analyze, tokenize and index MEDLINE document's title and abstract respectively. Moreover, MetaMap is employed to analyze the title and abstract text to map a set of associated CUIs which are indexed as well.

In the retrieval stage, shown as Step 5a, 5b in Figure 17, query text is analyzed by the same Lucene analyzer to extract query terms. MetaMap is used to map Metathesaurus CUIs from the query text. When the query's Original CUIs are mapped, we apply the Personalized PageRank algorithm to compute the PPV of that query described as Section

[2] http://lucene.apache.org/
[3] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart- stop-list/english.stop

4.4.3. Then we apply weight scheme in chapter 4.4.4 to construct final CUIs. Lastly, the free-text terms and Final CUIs (shown in Step 5c) are combined into a new expanded query for querying the Lucene index (Step 5d) in Figure 17.

4.5. Validation of Our Approach

4.5.1. Dataset

To evaluate the performance of our scheme, we compare the precision/recall of information retrieval under the same data set, the OHSUMED collection [9]. OHSUMED is a clinically-oriented MEDLINE subset, consisting of 348,566 documents covering all references from 270 medical journals over a five-year period (1987-1991). This dataset has been extensively utilized [10, 11, 13-15, 17] to carry-out BIR experiments. In creating the OHSUMED dataset novice physicians using MEDLINE generated 106 queries. Physicians were asked to provide a statement of information about their patients as well as their information need, or query. Each query was later replicated by four searchers, two physicians experienced in searching and two medical librarians. The results were assessed for relevance by a different group of physicians.

4.5.2. Experimental Design

Seven strategies are evaluated and compared in our experiment listed in Table 11.

- *Free-text*: Both title and abstract text of document and query text are analyzed and tokenized by Lucene's standard analyzer with enlarged stop-list and Porter stemmer.
- *Original CUIs*: Metathesaurus CUIs mapped by MetaMap tool and presented in four selected vocabularies.
- *Original CUIs + PRF*: it applies Pseudo Relevance Feedback (PRF) based query expansion on CUIs. The top 50 initially retrieved documents are collected, and the scores of the CUIs included in those documents are accumulated. Top ranked PRF CUIs are used to construct a new query.
- *(Original CUIs + PRF)* $\cap$ *Final CUIs*: the query expansion is based on the intersection between the PRF CUIs and Final CUIs. PRF scores are used.
- *Original CUIs* $\cup$ *Final CUIs*: the new expanded query includes Original CUIs in the first place; then it appends the top ranked PPV CUI candidates in the end, but skipping the already added Original CUIs. All CUIs in the final query are boosted by value b.
- *Final CUIs*: the new query is directly formed by the top ranked Final CUIs with boost value b. It's worth noting that Original CUIs are not guaranteed to be included in the new query.

Table 11: Seven index and retrieval strategies (*N/A: not applicable)

Retrieval Strategies	Document Representation		Query Representation	
	Vector 1	Vector 2	Vector 1	Vector 2
S1	Free-text	N/A	Free-text	N/A
S2	N/A	Original CUIs	N/A	Original CUIs

S3	Free-text	Original CUIs	Free-text	Original CUIs
S4	Free-text	Original CUIs	Free-text	Original CUIs + PRF
S5	Free-text	Original CUIs	Free-text	(Original CUIs + PRF) $\cap$ Final CUIs
S6	Free-text	Original CUIs	Free-text	Original CUIs $\cup$ Final CUIs
S7	Free-text	Original CUIs	Free-text	Final CUIs

Among the seven tested strategies, Strategies 1-4 repeat the work of previous studies and serve as a solid base line, and Strategies 5-7 apply our proposed method in query expansion in different ways.

4.5.3. Experimental Results

Following experiments use *Origin* ontology graph (built by four vocabularies) to compute personalized PageRank vector. Table 12 shows the *eleven points interpolated average precision* (11pt. avg. precision) at the 11 standard recall levels, *Mean Average Precision* (MAP), and *R-precision* [37]. *11-point interpolated average precision* is a traditional method to boil the precision-recall curve into eleven numerical values that the interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. To further simplify the performance of recall-precision, *Mean Average Precision* is widely used in TREC community providing a single-figure measure of quality across recall levels. Among evaluation measures, MAP has been shown to have especially good discrimination and stability. *R-precision* measures precisions at fixed low levels of retrieved results, such as 10 or 30 documents. All the three performance indicators can be

calculated by `trec_eval` tool[4] for the various retrieval strategies tested in this dissertation.

Table 12: Best performance of seven strategies

	S1	S2	S3	S4	S5	S6	S7
iprec_at_recall_0.00	0.7032	0.5594	0.7037	0.7029	0.6968	0.7226	0.7601
iprec_at_recall_0.10	0.5157	0.3637	0.5210	0.5309	0.5388	0.5509	0.5883
iprec_at_recall_0.20	0.4130	0.2728	0.4060	0.4345	0.4372	0.4283	0.4781
iprec_at_recall_0.30	0.3203	0.1960	0.3244	0.3479	0.3475	0.3358	0.3896
iprec_at_recall_0.40	0.2477	0.1389	0.2516	0.2863	0.2790	0.2614	0.3033
iprec_at_recall_0.50	0.2062	0.0883	0.1994	0.2393	0.2272	0.2121	0.2479
iprec_at_recall_0.60	0.1588	0.0566	0.1490	0.1827	0.1749	0.1601	0.1924
iprec_at_recall_0.70	0.1132	0.0357	0.0994	0.1349	0.1290	0.1120	0.1416
iprec_at_recall_0.80	0.0717	0.0219	0.0597	0.0850	0.0762	0.0675	0.0906
iprec_at_recall_0.90	0.0365	0.0119	0.0310	0.0408	0.0401	0.0330	0.0399
iprec_at_recall_1.00	0.0059	0.0008	0.0048	0.0063	0.0061	0.0047	0.0047
11pt. avg. precision	0.2538	0.1587	0.2500	0.2720	0.2684	0.2626	**0.2942**
MAP	0.2333	0.1366	0.2289	0.2530	0.2486	0.2415	**0.2704**
R-precision	0.2712	0.1810	0.2742	0.2907	0.2924	0.2840	**0.3060**

Table 13 presents the parameters used to achieve the best performance in different strategies. Table 14 shows the pairwise comparison between these strategies. A pair of strategies is compared by computing the percentage improvement achieved when using the stronger strategy over the weaker one. For example, row 2 column 3 indicates that S3 offers 57.5% improvement over S2.

[4] http://trec.nist.gov/trec_eval/

The result of S1-S3 conform to the observations in previous studies [10, 11] that strategy indexing both free-text and Metathesaurus information (S3) did not perform better than free-text indexing strategy (S1), and indexing restricted to Metathesaurus (S2) performed significantly worse than free-text strategy (S1). The pseudo relevance feedback strategy (S4) [13] improves the performance by 8.8% compare to the baseline S3.

Since S3 utilizes both free-text and Metathesaurus information and S4 applies additional query expansion, they serve as two solid base-line strategies to benchmark our proposed ontology graph based Strategies 5-7. S5 reconstructs the query by intersecting the set of S4 and Final CUIs which causes 1.3% drop (its term score uses PRF score rather than PPV weight). S6 avoids PRF and directly uses the PPV CUI candidates, but it keeps the original mapped CUIs; S6 is 5% better than S3, but 3.8% worse than S4. The best strategy S7 simply uses Final CUIs where part of Original CUIs may be excluded from the new query. To our surprise, S7's performance is significantly better than any other strategies where it improves 15.9%, 85.4%, 17.7% and 8.2% over baseline S1, S2, S3, S4 respectively. On average, S7 is 24.8% better than all other strategies.

Table 13: Parameters of best performance (*N/A: not applicable)

	S1	S2	S3	S4	S5	S6	S7
max retrieved #docs per query	100000	100000	100000	100000	100000	100000	100000
#docs for pseudo relevance feedback	N/A	N/A	N/A	50	50	N/A	N/A
#CUIs chosen for expanded query	N/A	N/A	N/A	5	15	25	15

boosting value b for expanded terms	N/A	N/A	N/A	0.4	0.75	0.7	0.8
α in Equation	N/A	N/A	N/A	N/A	N/A	0.1	0.1

Table 14: Pairwise comparison of retrieval strategies of 11pt. avg. precision

	S1 (0.2538)	S2 (0.1587)	S3 (0.2500)	S4 (0.2720)	S5 (0.2684)	S6 (0.2626)	S7 (0.2942)
S1 (0.2538)		-59.9%	-1.5%	7.2%	5.8%	3.5%	15.9%
S2 (0.1587)			57.5%	71.4%	69.1%	65.5%	85.4%
S3 (0.2500)				8.8%	7.4%	5.0%	17.7%
S4 (0.2720)					-1.3%	-3.8%	8.2%
S5 (0.2684)						-2.2%	9.6%
S6 (0.2626)							12.0%

4.5.4. Effectiveness Analysis

To effectively demonstrate the power of ontology graph based query expansion, we analyze the PPV CUIs generated from OHSUMED query. The details of query #10 "*Effectiveness of gallium therapy for hypercalcemia*" is presented in Table 15. MetaMap maps four Original CUIs for query #10: (C1280519: Effectiveness), (C0016980: Gallium), (C0039798: therapy), (C0020437: Hypercalcemia).

A close look at Table 15 leads us to believe that there are two key reasons why our proposed scheme performs better: (1) Ontology graph based query ranks specialized CUIs (Gallium, Hypercalcemia) much higher than generalized CUIs (effectiveness, therapy) because specialized CUI has a much larger IPF than generalized CUI. Thus, the ontology graph based query expansion has a less tendency to include those generalized

CUIs which may retrieve irrelevant noise documents. For query #10, only one Original CUI 'Gallium' is presented in the new query. (2) It successfully finds additional useful CUIs closely related to those valuable specialized CUIs (Gallium, Hypercalcemia). Rankings #1, #3, #8 are valuable CUI expansion for 'Gallium', and rankings #5, #6, #14, #15 are valuable CUI expansion for 'Hypercalcemia' in Table 15.

To demonstrate that using all Original CUIs can degrade the performance, we apply the same parameters set of S7 to S6 in Table 13. The evaluation result of S6 is: 11pt. avg. precision 0.2597, MAP 0.2386, R-precision 0.2831. The result shows that the generalized terms in Original CUIs can degrade the performance as much as 13.3% in 11pt. avg. precision.

Table 15: Details of PPV final weights of OHSUMED Query #10 "*Effectiveness of gallium therapy for hypercalcemia*" (asterisk * indicates Original CUIs)

Rank	PPV CUI	Final Weight	Init. PPV Score s_i	IPF	Concept Name
1	C0202390	5.8315	0.0006	12.341	Gallium measurement
*2	C0016980	5.7205	0.0756	7.4059	Gallium
3	C0061005	5.5911	0.0006	11.8302	gallium arsenide
4	C0150195	5.4806	0.0008	11.2424	Electrolyte management: hypercalcemia
5	C1833372	5.4740	0.0007	11.2424	Familial benign hypercalcemia, type 3
6	C0682902	5.3936	0.0006	11.2424	boron group elements
7	C0878684	5.3856	0.0008	11.0417	SHORT syndrome
8	C0061008	5.2749	0.0011	10.395	gallium nitrate
9	C0268478	5.2404	0.0008	10.7315	Blue diaper syndrome
10	C0033597	5.1809	0.0011	10.2207	Protactinium
11	C0005124	5.1423	0.0011	10.1437	Berkelium
12	C0015853	5.1061	0.0011	10.0723	Fermium
13	C0025275	5.0723	0.0011	10.0056	Mendelevium
14	C0271851	4.9913	0.0008	10.1437	Hypercalcemia due to sarcoidosis
15	C0271850	4.9901	0.0008	10.1437	Hypercalcemia due to granulomatous disease

N			N		N
*27	C0020437	4.8283	0.0732	6.2714	Hypercalcemia
N			N		N
*275	C1280519	1.7453	0.0782	2.2518	Effectiveness
N			N		N
*362	C0039798	0.5907	0.0733	0.7672	Therapy

4.5.5. Multi-Vocabularies of Ontology Graphs

Recall the above experiments are performed on Origin ontology graph (built by four vocabularies). Now, we want to study the effectiveness by enlarging the ontology graph with more vocabularies. Thus, we had performed a series of additional evaluations on *Medium* and *Large* ontology graphs shown in Figure 19. *Medium* ontology graph uses eight vocabularies and *Large* ontology graph uses eleven, shown in Table 10. We use them to re-compute the PPVs from 106 queries and 348K documents. Finally, the 11 point average precision values are calculated with the same parameters set as S7 in Table 13. We vary the size of PPV CUIs before re-ranking by weighted scheme. It shows that small size of CUIs (<150) degrades the performance greatly. Large size of CUIs (>500) doesn't play a role in the final value. The size between 200 and 250 shows the best result.

Figure 19 also shows that the Origin ontology graph still performs the best, and *Large* ontology graph's performance is better than *Medium* ontology graph. It implies that increasing the number of ontologies may not improve the overall performance. Further experiments are required to identify which vocabulary causes the performance degradation.

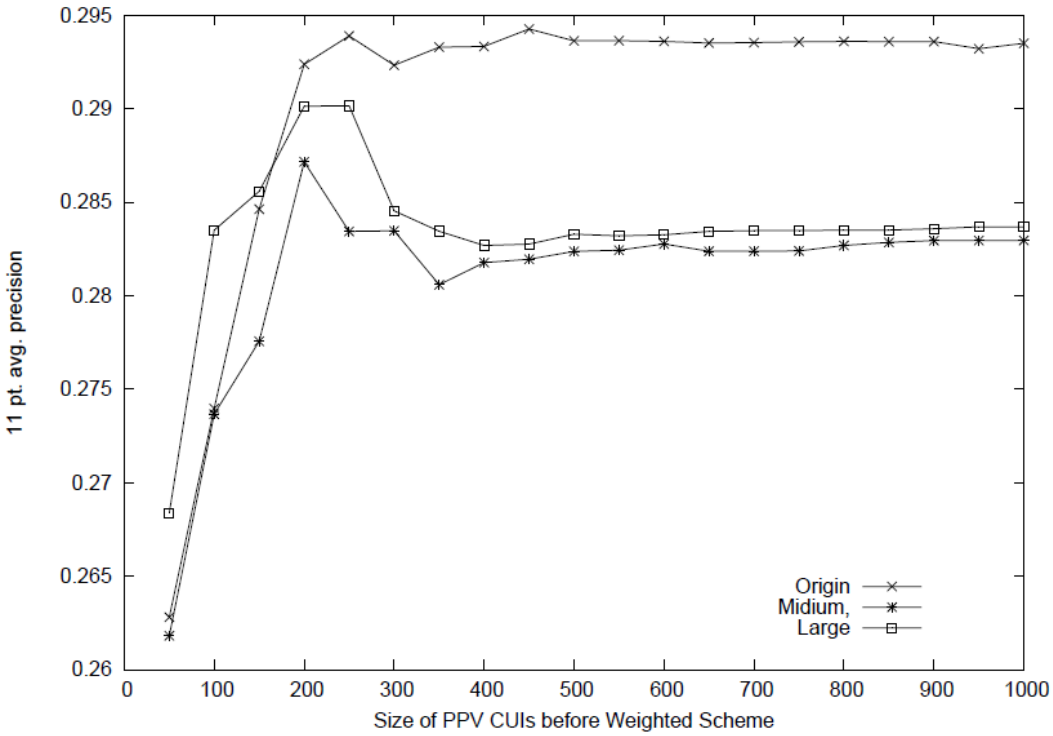


Figure 19: 11pt. avg. precision values using three different ontology graphs (with 4, 8 and 11 vocabularies respectively) selecting various size of PPV CUIs before we re-rank those CUIs by Weighted Scheme.

4.6. Summary

We have proposed a new ontology graph based query expansion scheme for MEDLINE. MeSH and three other controlled vocabularies from Metathesaurus are used to construct the graph. MetaMapped biomedical concepts are used to find semantically related counterparts by running Personalized PageRank algorithm on the graph. A carefully designed weight scheme is applied to select top biomedical concept candidates for query expansion. Experiments show that the best ontology graph based query expansion S7 surpasses the results of pseudo relevance feedback based query expansion

S4, no query expansion S3, and all other strategies by 8.2%, 17.7% and 24.8% on average in 11pt. interpolated average precision. We also identify that the generalized biomedical concept is one of the reasons for performance degradation.

Chapter 5

Hybrid Query Expansion Assisted by WEST

5.1. Background

In this chapter, we will apply the Weighted Edge Similarity (WEST) method from chapter 3 into our previously successful PPV query expansion approach for biomedical information retrieval from chapter 4.

Directly applying word semantic similarity into query expansion isn't an easy task. Voorhees [87] showed that an automatic procedure of query expansion based on the WordNet synonym sets can degrade retrieval performance. His experiments showed that the query expansion technique makes little difference in retrieval effectiveness if the original queries are relatively complete descriptions of the information being sought even when the concepts to be expanded are selected by hands; while less well developed queries can be significantly improved by expansion of hand-chosen concepts.

Jalali [88, 89] applied Li's similarity method [90] on MeSH tree ontology by computing the word similarity between the original query terms and pseudo relevance

feedback terms. A threshold of 0.7 is used to cut-off low similar terms in the pseudo relevance feedback procedures in his approach.

5.2. Hierarchy of Ontology Graph

To apply our WEST algorithm, there are two prerequisite conditions to satisfy:

(1) Whether there exists a suitable underlying ontology structure?

(2) Whether the hierarchy of ontology structure can be explored and Least Common Ancestor can be computed?

Luckily, after carefully studying the Metathesaurus ontology, we find that both prerequisite conditions can be fulfilled by using multiple biomedical ontologies derived from Metathesaurus.

First, since we are working on the biomedical data, the underlying ontology has to be changed from WordNet to the Metathesaurus ontologies which were built in chapter 4. We choose to use the *Origin* ontology graph of four vocabularies (MSH, SNOMEDCT, AOD, CSP) for its simplicity and effectiveness in the following experiments.

Second, the hierarchy of the ontology has been constructed in the “Computable” Hierarchies (MRHIER) table of UMLS Metathesaurus. The MRHIER table of the *Origin* ontology graph was constructed by the four vocabularies with 6,876,273 total records of which 278,085 distinct CUIs.

The MRHIER table has two important attributes: AUI and PTR. AUI is short for Atom Unique Identifiers [91] which is the basic building blocks or "atoms" from which the Metathesaurus is constructed from each of the source vocabularies. Every occurrence

of a string in each source vocabulary is assigned a unique atom identifier or AUI. If exactly the same string appears multiple times in the same vocabulary, for example as an alternate name for different concepts, a unique AUI is assigned for each occurrence. AUI contain the letter A followed by seven numbers. The abbreviation for the source that contributed each string is noted in parentheses after the string.

Table 16: MRCONSO of CUI C0016980 “Gallium”

AUI	SAB	STR
A0014095	MSH	Gallium
A2877777	SNOMEDCT	Gallium
A0014094	MSH	Gallium
A0479659	CSP	gallium
A0479658	AOD	gallium
A4781508	SNOMEDCT	Gallium, NOS
A1961887	CSP	Ga element
A3471456	SNOMEDCT	Gallium (substance)

Table 17: MRHIER of CUI C0016980 “Gallium”

AUI	SAB	PAUI	PTR
A0014094	MSH	A0743535	A0434168.A2367943.A18456972.A0135374.A0135450.A0053536.A0743535
A0014094	MSH	A0743535	A0434168.A2367943.A18456972.A0135374.A0135450.A0085365.A0743535
A0479658	AOD	A1388564	A1386158.A1389303.A1389283.A1392037.A1388564
A0479659	CSP	A1195034	A0398472.A0318590.A0318854.A0483678.A1195034
A2877777	SNOMEDCT	A3471460	A3684559.A3206010.A16967690.A3347798.A3559706.A3471460
A2877777	SNOMEDCT	A3471460	A3684559.A3206010.A3738095.A3347798.A3559706.A3471460

PTR denotes for “Path to Top or Root” of the hierarchical context. The PTR is a string composed of AUI separated by periods, each AUI representing a node in the Metathesaurus hierarchy. The PTR and the AUI were concatenated to produce a

Hierarchical Unique Identifier (HUI) locating the given record in the Metathesaurus hierarchy [92].

Other attributes of MRHIER includes SAB and PAUI. SAB is short for “Source Abbreviation” which records which vocabulary it is stored. The PAUI shows the direct parent of that CUI. There are three CUIs in our version of the MRHIER which don’t have PTR values: Medical Subject Headings (C1135584/A0434168), CRISP Thesaurus (C1140093/A0398472), Alcohol and Other Drug Thesaurus (C1140162/A1386158).

To better illustrate the hierarchy provided by Metathesaurus, we re-use the OHSUMED query #10 “Effectiveness of gallium therapy for hypercalcemia” from chapter 4.4.4. The term “Gallium” is corresponding to CUI C0016980 in MRCONSO in Table 16 and MRHIER in Table 17.

Figure 20 shows the hierarchy of the ontology graph between CUI pair <“gallium”, “gallium nitrate”>, < “gallium”, “fermium”> and < “gallium”, “berkelium”>. The AUI specific level (SpecLev) of the hierarchy is shown in the figure which is used to compute the weighted length as well as the similarity value of a pair.

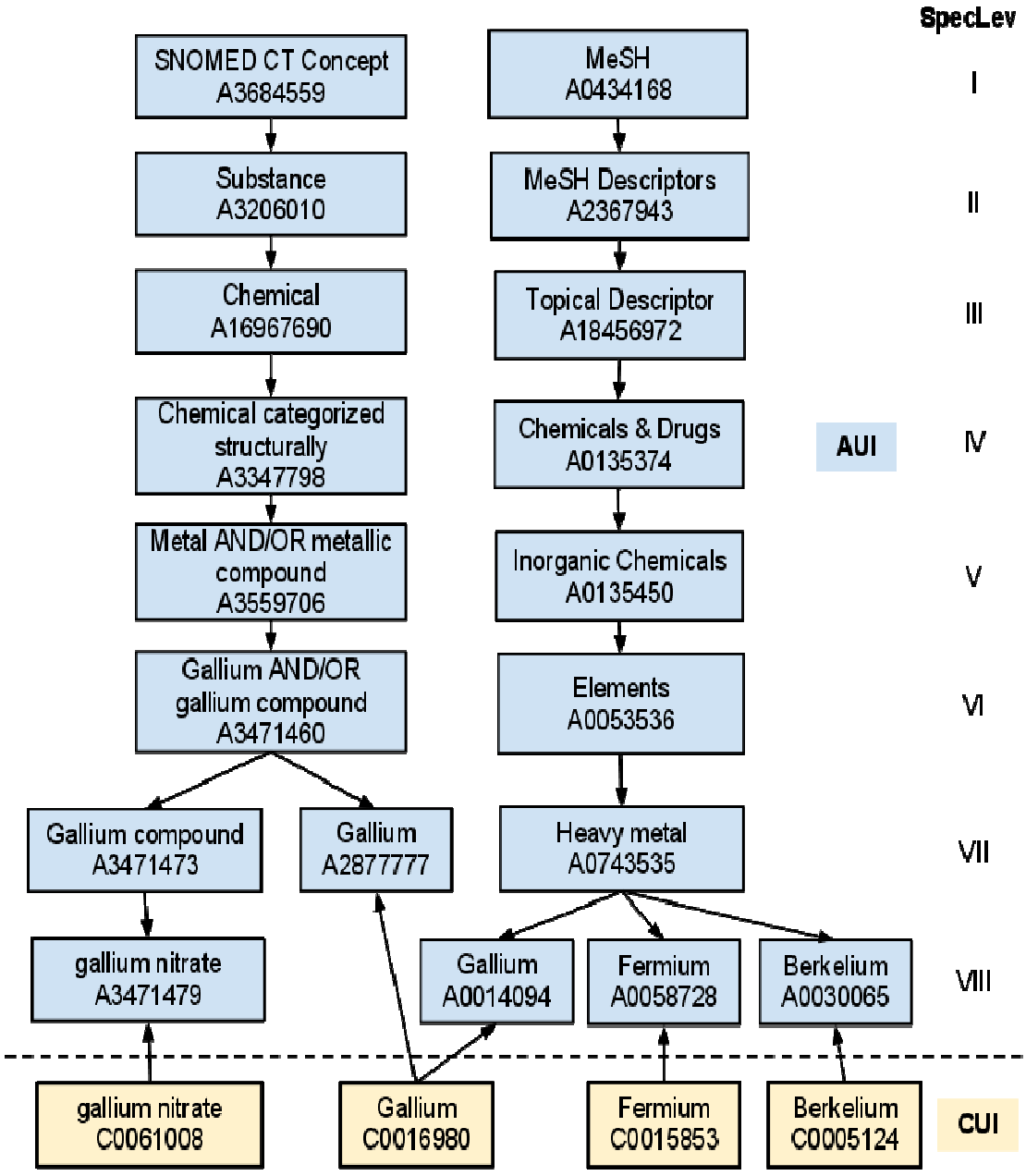


Figure 20: Hierarchy of “gallium nitrate”, “fermium”, “berkelium” and “gallium”

5.3. Weighted Edge Similarity Assisted Query Expansion

In this section, we applied our Weighted Edge Similarity (WEST) algorithm on the ontology graph to compute the semantic similarity of the Final CUIs and the Original CUIs.

The motivation of applying semantic similarity to screen Final CUIs is to further considering the generalization and specification of the Final CUIs. Since the personalized PageRank algorithm only considers the in-link relationship and we use the weight scheme to filter those CUIs with high document frequency. However, it doesn't consider the Final CUIs' relationship in the way whether the expanded CUI is more general or more specific of the Original CUIs. By applying the Weighted Edge Similarity algorithm, we are able to filter those more general expanded CUIs and keep those more specific expanded CUIs into the final expanded query.

The WEST algorithm is applied in Step 5c in Figure 21 noted in red color. The rest of the flow chart is the same of the personalized PageRank (PPV) based Query Expansion.

In the Step 5c, we evaluate the top *K* Final CUIs and compute the semantic similarity of each Final CUI with all the Original CUIs and keep the highest similarity value. A heuristic similarity threshold is set according to the decreasing rate α value of WEST. If a Final CUI's highest similarity value is lower than the threshold, then that CUI will be skipped in the final expanded query.

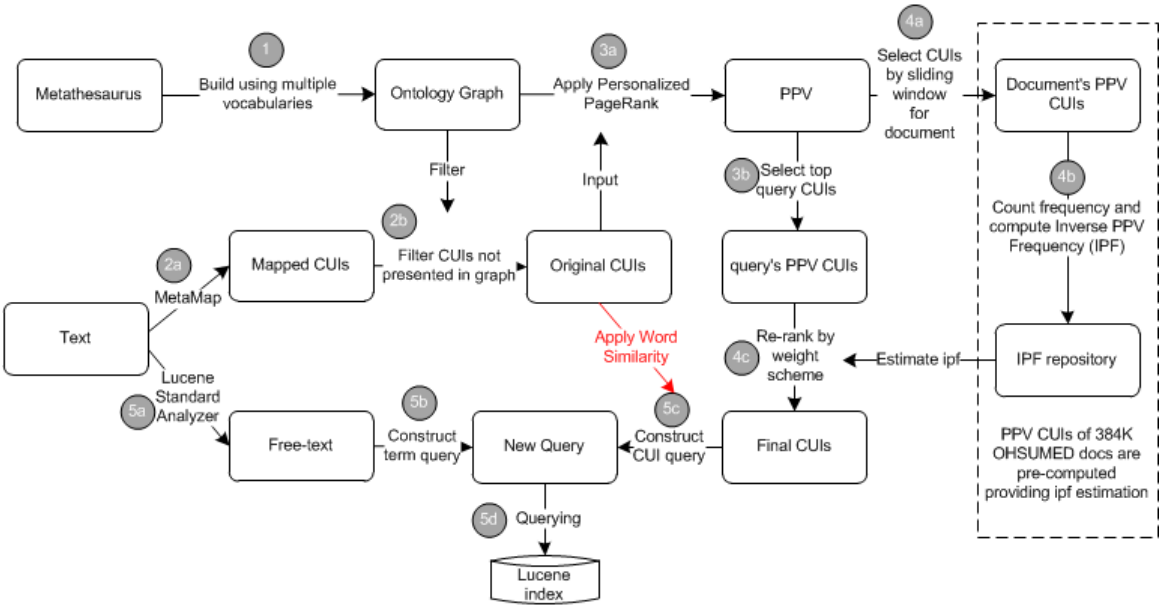


Figure 21: Flow chart of applying Weighted Edge Similarity (WEST) algorithm

5.4. Validation of Our Final Query Expansion Approach

To validate our approach applying WEST algorithm into the personalized PageRank based query expansion, we implement the WEST algorithm on the Metathesaurus ontology graph. Three additional experiments are performed to evaluate the performance gain on the original PPV based query expansion.

Table 18: Three WEST assisted index and retrieval strategies

Retrieval Strategies	Document Representation		Query Representation	
	Vector 1	Vector 2	Vector 1	Vector 2
S7	**Free-text**	**Original CUIs**	**Free-text**	**Final CUIs**
S8	**Free-text**	**Original CUIs**	**Free-text**	**WEST(sech) Final CUIs**
S9	**Free-text**	**Original CUIs**	**Free-text**	**WEST(tanhc) Final CUIs**
S10	**Free-text**	**Original CUIs**	**Free-text**	**WEST(sech*tanhc) Final CUIs**

The best strategy of personalized PageRank algorithm S7 is used as the baseline. We evaluate the three hyperbolic transfer functions described in chapter 3.5.2 in Table 18. S8 uses *sech* function, S9 uses *tanhc* function and S10 applies *sech*tanhc* which again

showed the best performance in our WordNet experiments. In this WEST assisted query expansion experiment, a heuristic threshold value is set to 0.30 which is close to our previous work [18, 19], and the decreasing rate of WEST algorithm $\alpha = 0.8$ for all three new strategies. The experimental result is shown in Table 19.

Table 19: Performance of WEST assisted hybrid query expansions

	S7	S8	S9	S10
WEST transfer function	N/A	sech	tanhc	sech* tanhc
iprec_at_recall_0.00	0.7601	0.7475	0.7344	0.7775
iprec_at_recall_0.10	0.5883	0.5940	0.5787	0.6034
iprec_at_recall_0.20	0.4781	0.4889	0.4761	0.4912
iprec_at_recall_0.30	0.3896	0.4068	0.3891	0.4092
iprec_at_recall_0.40	0.3033	0.3283	0.3144	0.3291
iprec_at_recall_0.50	0.2479	0.2596	0.2621	0.2741
iprec_at_recall_0.60	0.1924	0.2025	0.2005	0.2170
iprec_at_recall_0.70	0.1416	0.1494	0.1443	0.1657
iprec_at_recall_0.80	0.0906	0.0933	0.0892	0.0941
iprec_at_recall_0.90	0.0399	0.0386	0.0435	0.0467
iprec_at_recall_1.00	0.0047	0.0053	0.0061	0.0073
11pt. avg. precision	0.2942	0.3013	0.2944	**0.3105**
MAP	0.2704	0.2841	0.2716	**0.2857**
R-precision	0.3060	0.3176	0.3086	**0.3252**

The experiment shows that all of three new strategies improve the personalized PageRank query expansion. Among three strategies, the best strategy S10 applying both sech and tanhc as the transfer function improves the eleven point average precision by 5.54% comparing to S7 and 22.34% to S1.

5.5 Discussion

We use weighted edge similarity algorithm to assist word expansion by further filtering low similarity terms from the expanded terms generated by Personalized PageRank algorithm. Experiments show that all three strategies S8-S10 with WEST improve the search performance comparing to those method without applying similarity filtering.

The reason for performance improvement is due to the removal of general concept and kept of specific concept. Personalized PageRank algorithm selects the concepts which best matches the query context; while the weighted scheme re-weights the entire rank so that general concepts are ranked lower and specific concepts are ranked higher. WEST similarity further filter those general concepts based on its specific level in the ontology that terms with low specific level (more general) are removed from the expansion list.

Chapter 6

G-Bean: A Graph-based Biomedical Search Engine

6.1. Overview

We have implemented an interactive Graph-based Biomedical Search Engine (G-Bean) using our proposed ontology graph query expansion algorithm. The online system accepts any medical related user query and processes them with expanded medical query to search for the whole MEDLINE database.

6.2. Architectural Design

6.2.1. MEDLINE Dataset

It is not trivial and fairly important to collect the entire corpus of MEDLINE records as well as MetaMapping the entire MEDLINE text contents. Our first trial is to manually create a Python script to crawl the MEDLINE records from NLM's *entrez* portal [93]. It spends us more than 10 days to crawl 14M records.

However, at the same time, we found that NLM has already built a package called Medline/PubMed Baseline which contains the entire MEDLINE. More importantly, the Medline/PubMed Baseline has already applied MetaMap to the whole MEDLINE data parsing these citations and get corresponding CUIs for every citation [94]. According to the description, the entire MEDLINE corpus of 19,569,568 citations was created on November 19, 2011. It was processed (by shell command *metamap10 –Z 1011 –qE*) between January 26, 2011 and February 16, 2011 through the MetaMap program generating MetaMap Machine Output formatted results for each of the citations. The results are now available via the link [95]. The compressed downloadable data requires 129.9GB disk space.

Thanks to the NLM's pre-processed MEDLINE citation data which saved us more than 20 days of work, we apply our information retrieval model to index the MEDLINE as well as its MetaMap processed CUIs as shown in Chapter 4.3.5. However, building an index for such a large scale data is challenge even using Lucene library. In real index phrase, we repeated several times trying to index 20M citations and failed due to the Java virtual memory space is not enough. We finally succeed our approach by optimizing the Lucene index at every 50 input files (total 653 files) and setting the Java virtual machine's memory by *-Xms4096m -Xmx4096m*.

At first, we process the whole MEDLINE citations by indexing its title and abstract processed by porter stemmer and filtered by MIT stop-list[5]. The MetaMapped

[5] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart- stop-list/english.stop

CUIs are also indexed for our proposed query expansion. The whole indexed data requires 25.8GB disk spaces.

However, when we evaluate the index we created as above, we find that using porter stemmer and stop-list is not a good option for biomedical document indexing. The reason is that some biomedical special terms will be removed during indexing and searching phase. For example, Gene Ontology is short for (GO) which is in the stop-list. When we search the term G-SESAME, it returns documents about sesame which is not what we want.

In order to solve this problem, we re-index the entire document corpus simply using white space to separate each term. We do not use porter stemmer, stop-list or distinguishing capitalized letter in the second round indexing. The re-indexed data requires 39GB disk space and takes 18.2 hours to index.

6.2.2. Architecture

Since we are using the Java version Lucene library underlying our query expansion implementation, we choose to implement the online system using Client-Server architecture powered by Java Servlet Pages. The front-end is written by Java Servlet Pages (JSP) and the back-end is supported by our ontology graphed assisted hybrid query expansion system. The detailed architecture is shown in Figure 22.

As shown in Figure 22, the front-end is composed by HTML and JSP codes which are directly displayed to users around the world. When the user’s query is passed to the back-end system, the original query is parsed via Porter stemmer and filtered by the MIT

stop-list. The MetaMap program searches and generates the corresponding CUIs recorded in the Metathesaurus. The CUIs will be expanded by running the personalized PageRank algorithm on the ontology graph at first. Then, the expanded CUIs will be filtered by computing the semantic similarity between the expanded CUIs and the Origin CUIs. The filtered Final CUIs with the original text phrases are composed together as the Hybrid Final Query to search our local MEDLINE indexes.

Currently, the proposed G-Bean search engine is deployed on web server *Tomcat 6.0* using Ubuntu 11.04 as the operating system. The current version of web application system is at *http://bioir.cs.clemson.edu:8080/BioIRWeb/index.jsp*.

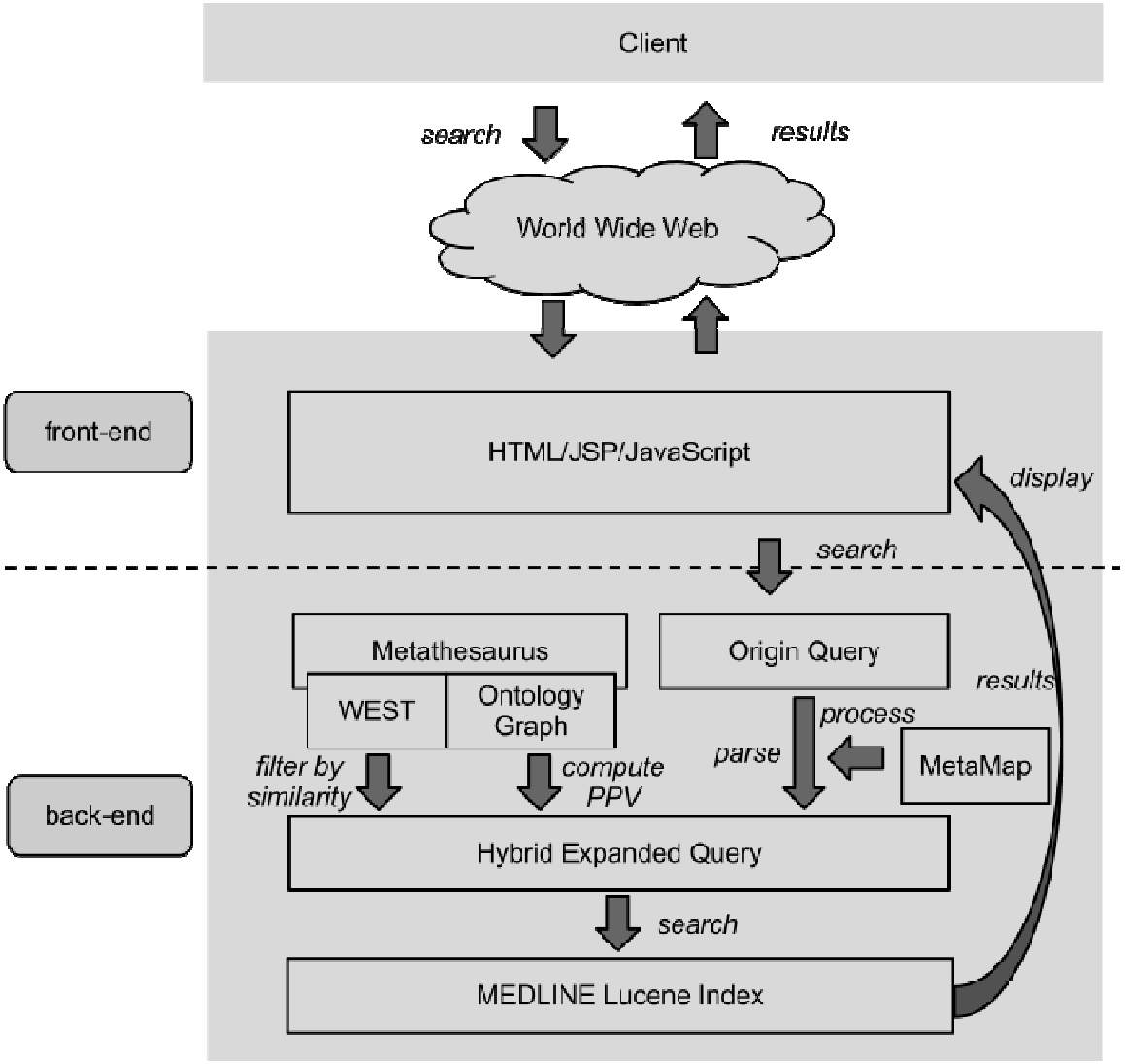


Figure 22: Architecture of G-Bean

6.3. Usage

The interface of the website is shown Figure 23 where user can query any biomedical terms and G-Bean returns a list of biomedical documents from MEDLINE database. The current URL of the website is at

http://bioir.cs.clemson.edu:8080/BioIRWeb. Click the title of any listed item will link to the original item in the PubMed online database in Figure 24.

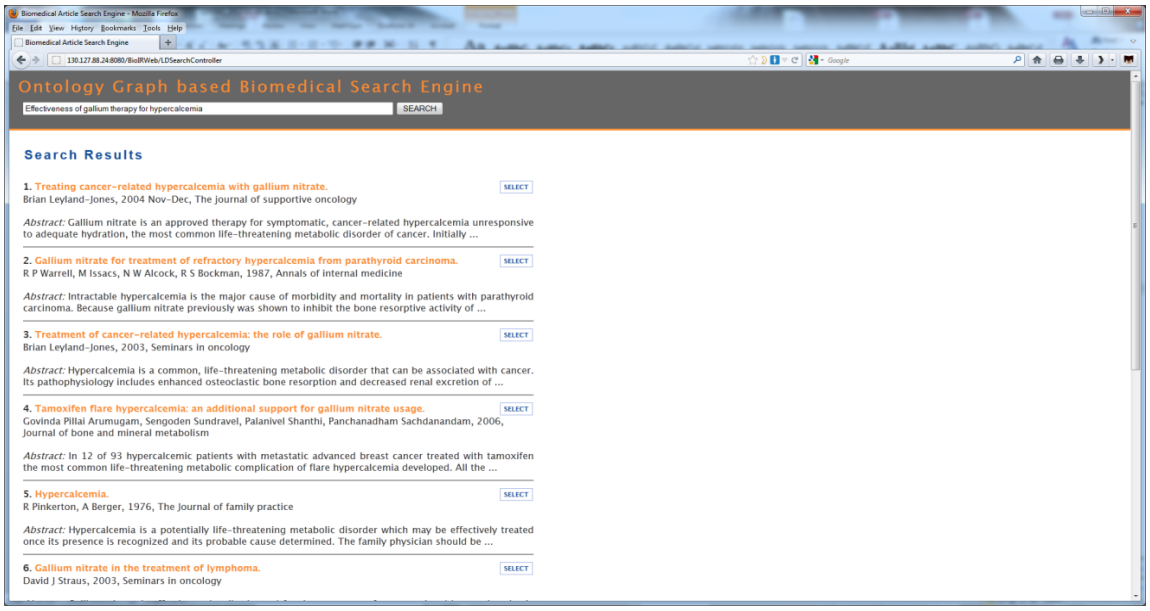


Figure 23: Biomedical information retrieval website

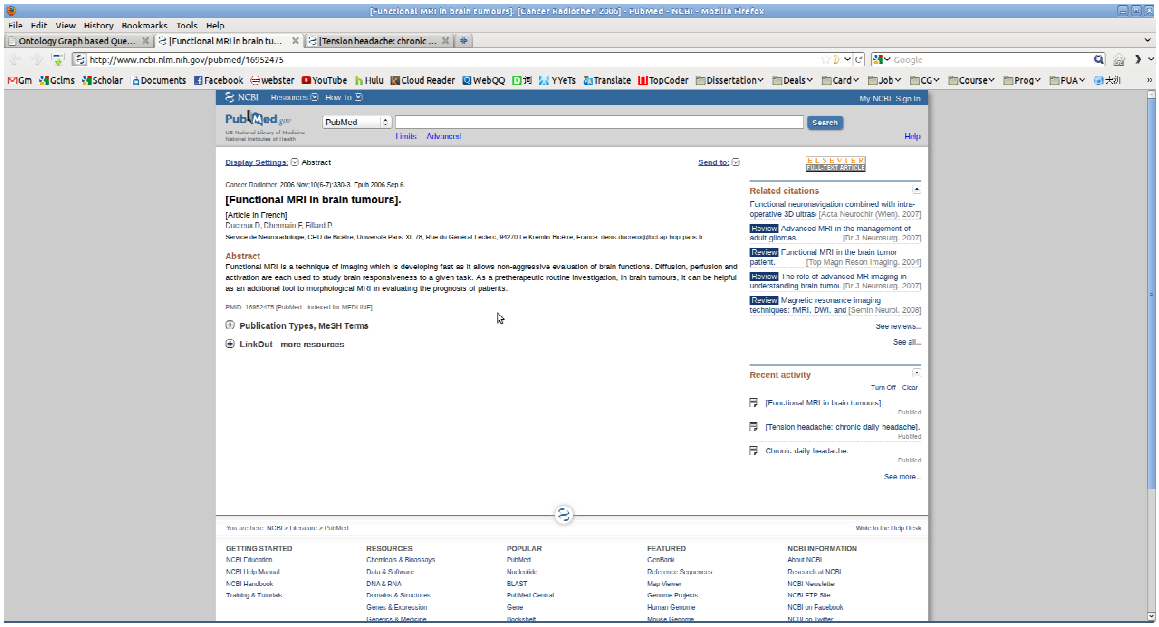


Figure 24: Selected article is linked to PubMed database

One feature of our website is that the user can select his interested article and find its related articles. The selected article is displayed in the middle column and the related articles are shown in the right column. As shown in Figure 25, the user adds “Treatment of cancer-related hypercalcemia the role of gallium nitrate” into the Selected Articles, and the right bottom articles shows top relevant articles to that user selected article.

User can select multiple articles and add them into Selected Articles in the middle column; while Related Articles in the right column includes the related articles for each selected article. All the related articles are re-sorted by its matched score. As shown in Figure 26, the user selects one more article “Gallium nitrate in the treatment of lymphoma”.

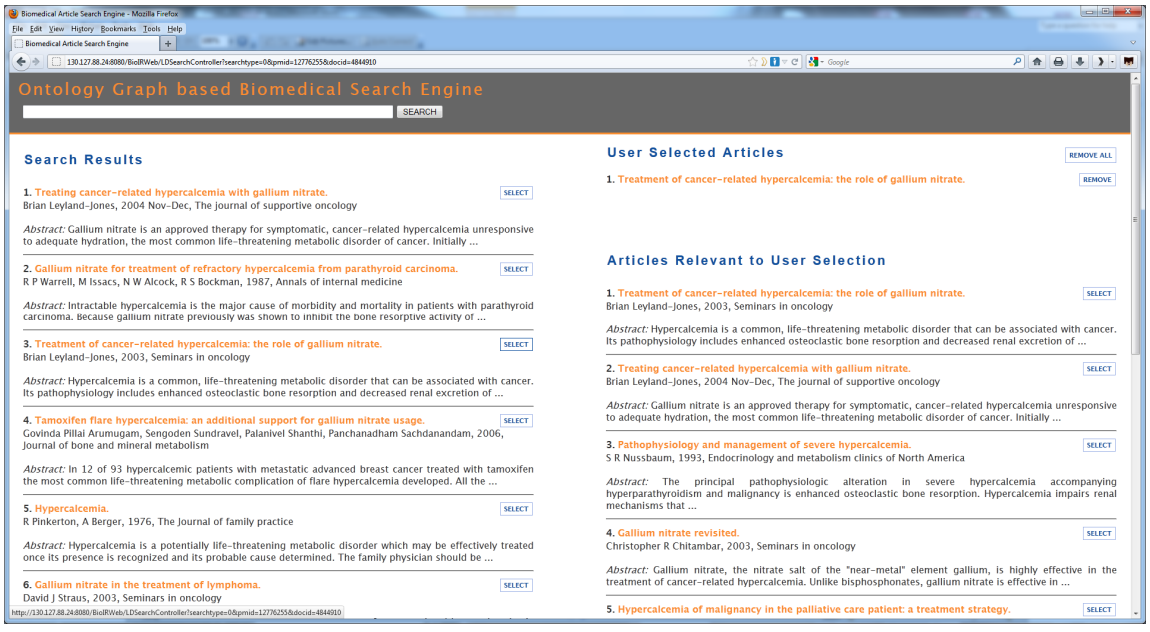


Figure 25: User selects article from search results

The user can change the query but keep the contents in the Selected Articles and Related Articles in order to select additional articles to the middle and right columns.

Figure 27 shows the corresponding selected result in the right columns when multiple articles are selected.

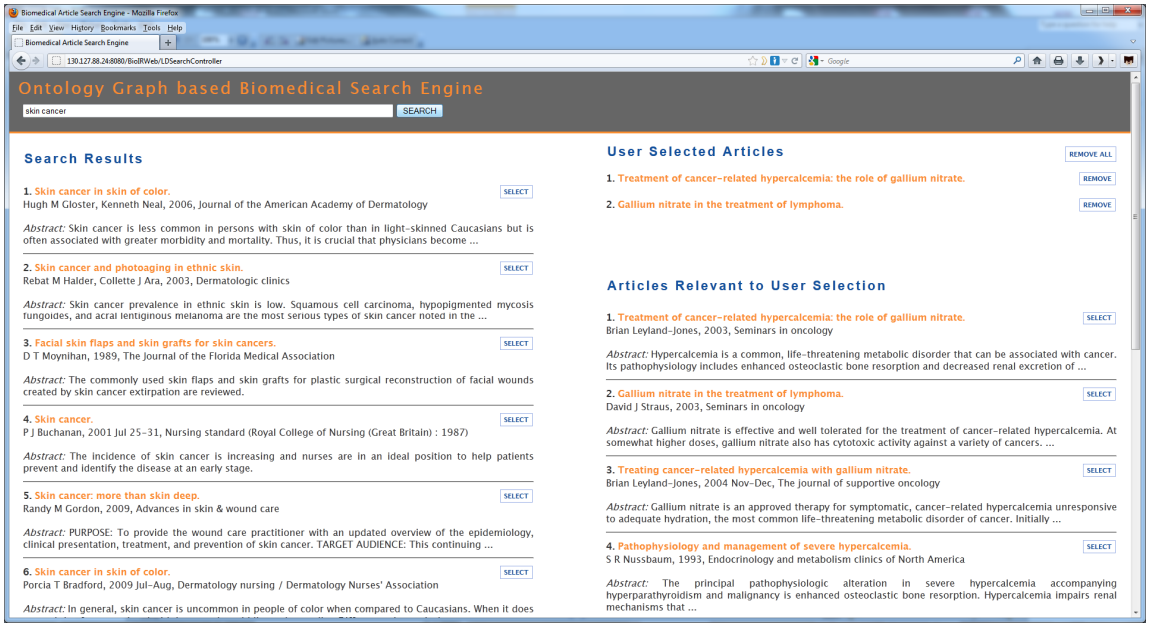


Figure 26: User selects additional article

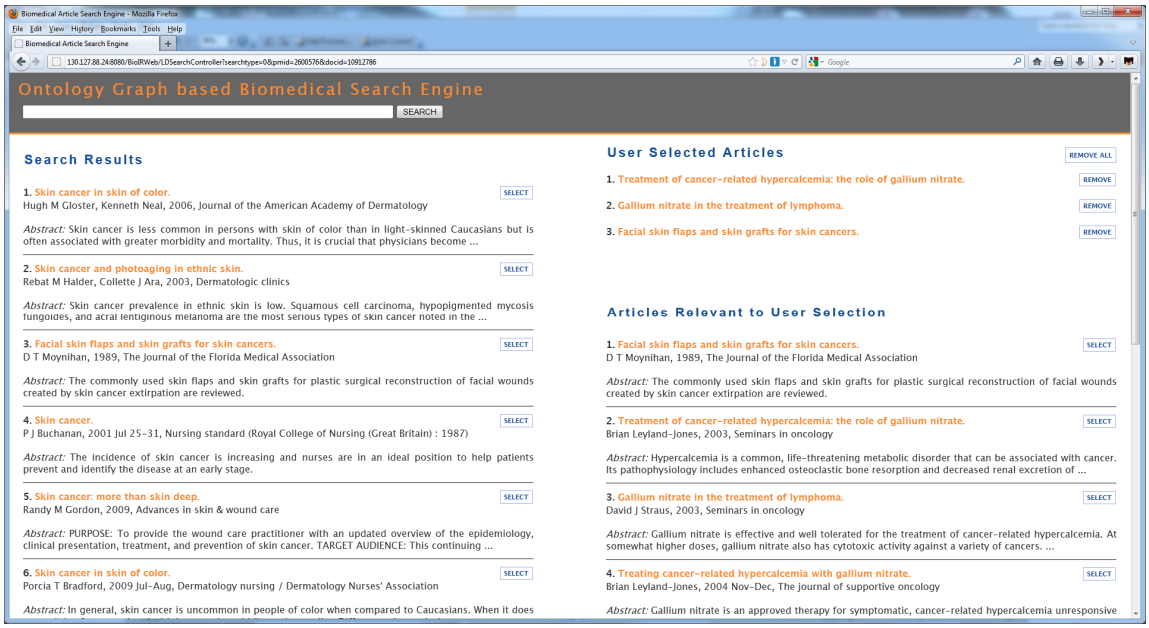


Figure 27: Change the search keywords to “skin cancer” and select additional article

6.4. Evaluation

Objective evaluation was shown in the previous chapters. In this section, we show our subjective evaluation comparing G-Bean and PubMed.

To evaluate the performance of our Graph based Biomedical Search Engine (G-Bean), we have used the 106 queries from OHSUMED dataset to search the entire 20 million MEDLINE citations. The search results were compared with the results returned by PubMed interface. An expert in biomedical sciences carefully examined the results returned by both search engines. Surprisingly, the expert felt that G-Bean returned better search results in 79 of these queries while both search engines returned good search results on other 27 queries. This evaluation further confirms the superiority of G-Bean biomedical search engine. It is worth-noting that PubMed system fails to return any results on several queries such as #7, #52, and #101.

From the biomedical expert's judgment, we find that if the query is composed of MeSH terms, both systems perform well. However, if the query cannot be parsed into MeSH terms, the PubMed usually doesn't return desired results and our system outperforms PubMed in most of the case. Besides, the PubMed system frequently matches items simply related to general terms such as "therapy" and "effective" which decrease the precision and degrade the performance. To sum up, our G-Bean system outperforms the original PubMed's search and it is more convenient for user to perform efficient and effective search in biomedical area.

Table 20 shows the OHSUMED Query #11 “review article on cholesterol emboli” where the term “*cholesterol emboli”* is not in the MeSH ontology. Thus, only #3 from

PubMed is related to the user query. However, our G-Bean is able to automatically mapping cholesterol emboli into its related CUI C0149649 which gives us a better result in our biomedical search engine that all the top 5 results are related to the user's query.

Table 20: Top 5 in OHSUMED Query #11 "review article on cholesterol emboli"

	PubMed	**G-Bean**
1	Pitfall in nephrology: contrast nephropathy has to be differentiated from renal damage due to atheroembolic disease.	Cutaneous cholesterol emboli (author's transl).
2	Objectives of teaching direct ophthalmoscopy to medical students.	Spinal cord infarction due to cholesterol emboli complicating intra-aortic balloon pumping (case report and review of the literature).
3	Cholesterol embolization syndrome.	Multiple cholesterol emboli syndrome. Bowel infarction after retrograde angiography.
4	Models of preventable disease: contrast-induced nephropathy and cardiac surgery-associated acute kidney injury.	Cholesterol emboli after cardiac catheterization. Eight cases and a review of the literature.
5	Subcutaneous thrombotic vasculopathy syndrome: an ominous condition reminiscent of calciphylaxis: calciphylaxis sine calcifications?	Multiple cholesterol emboli syndrome.

Table 21 shows top 5 articles retrieved by OHSUMED Query #19 "use of beta-blockers for thyrotoxicosis during pregnancy" using the two search engine. Only G-Bean is able to retrieve articles related to "beta-blockers" while the PubMed retrieved none articles related to beta-blockers.

Table 21: Top 5 in OHSUMED Query #19 “beta-blockers for thyrotoxicosis”

	PubMed	G-Bean
1	Therapy of hyperthyroidism in pregnancy and breastfeeding.	Treatment of thyrotoxicosis during pregnancy with propranolol.
2	Hyperthyroidism and other causes of thyrotoxicosis: management guidelines of the American Thyroid Association and American Association of Clinical Endocrinologists.	Oral beta-blockers for mild to moderate hypertension during pregnancy.
3	[Severe circulatory insufficiency in a patient with neonatal hyperthyroidism].	Evaluation of thyrotoxicosis during pregnancy with color flow Doppler sonography.
4	Molar pregnancy-induced thyroid storm.	Oral beta-blockers for mild to moderate hypertension during pregnancy.
5	Total intravenous anesthesia for evacuation of a hydatidiform mole and termination of pregnancy in a patient with thyrotoxicosis.	Transient post-operative thyrotoxicosis after parathyroidectomy.

The entire 106 OHSUMED queries, its top 5 results from both system and the biomedical expert’s opinions are presented in website: http://bioir.cs.clemson.edu/SearchEngineEvaluation/evaluation.php.

Several OHSUMED queries such as Query #23 “spontaneous unilateral galactorrhea, differential diagnosis and workup” and Query #30 don’t get results in PubMed while our search engine returns good results.

Based on these subjective evaluations, G-Bean is more stable and effective comparing to PubMed search, especially when user’s query contain terms which are not MeSH terminology.

Chapter 7

Conclusion

7.1. Contribution Summary

We have proposed an enhanced search engine for the biomedical research community to facilitate effective searches via a hybrid query expansion approach on biomedical ontology graph. The biomedical ontology graph can be constructed by any number of existing biomedical vocabularies in Metathesaurus which provides the possibility of customized search for different users. Two different but related methods exploring the ontology graph are studied and evaluated to construct an expanded query to search the MEDLINE Lucene index. Both of the methods are proved to be effective in increasing the recall-precision performance. To sum up, our contributions are ten-folds as listed below:

(1) Our proposed query expansion algorithm is conceptually novel and very different from previous query expansion methods in information retrieval as of our knowledge.

(2) Unlike most of the previous ontology based studies which utilize only MeSH as their solo ontology, our method can employ multiple controlled vocabularies from Metathesaurus for indexing and searching.

(3) The application of multiple vocabularies provides the possibility for users to customize their specialized search. A gene scientist can create the ontology using GO vocabulary to expand the query specifically to Gene Ontology.

(4) We have designed a systematic method to eliminate the mapped generalized biomedical concepts and populate closely related specialized concepts resulting in significant increase in the relevance of retrieval results.

(5) Our experimental analysis showed that eliminating generalized biomedical concepts in the search query may greatly improve the recall-precision performance.

(6) We demonstrate that query expansion based on ontology graph is more stable than that based on pseudo relevance feedback because sorting the retrieved documents by relevance is found to be often inaccurate.

(7) We made an important observation that humans are more sensitive to the word semantic difference caused by the categorization than by specification. In another word, people view word pair separated by specification more similar than those separated by categorization.

(8) Our WEST semantic similarity algorithm performs well on both WordNet and multiple ontologies generated from Metathesaurus.

(9) We explore two different yet effective approaches to take advantages of the multiple biomedical ontologies into bioinformatics information retrieval.

(10) The two approaches are successfully combined and the hybrid approach has achieved best performance in our experiments.

7.2. Future work

7.2.1. Further Evaluation of Multiple Ontologies

We explore the multi-vocabularies of ontology graph construction in Chapter 3.4.5. The *Origin* version with four vocabularies was increased with additional vocabularies to construct *Medium* version (8 vocabularies) and *Large* version (11 vocabularies) ontology graph. However, both Medium and Large version don't perform better than the Origin version while the Large version performs better than the Medium version. This implies that the introduction of certain ontology might impair the overall retrieval performance. A further evaluation of the relationship between the retrieval performance and the combination of multiple ontologies can be studied.

7.2.2. Speed-up the Personalized PageRank Computation

Currently, we compute the personalized PageRank vector on the fly during the query expansion construction phrase. The PPV computation for each query might take one to several seconds which is based on the size of the ontology graph. However, this process can be accelerated with several existing methods. One outstanding solution is

proposed by Jeh and Widom called Scaled Personalization in [96]. The authors developed an approach to compute PPV as a solution of a linear combination of a set of basic PPVs. For a given teleport vector v, the personalized PageRank equation can be deduced into Equation (22):

$$x = Ax = cP^T x + (1-c)v, 0 < c < 1 \tag{22}$$

where the PPV x relates to user-specified bookmarks with weights represented in v [83]. The author Haveliwala proposed the Linearity Theorem to encode PPV into shared components:

Linearity Theorem. *The solution to a linear combination of preference vectors* v_1 *and* v_2 *is the same linear combination of the corresponding PPV's teleport vector* x_1 *and* x_2, *for any constants* $\alpha_1, \alpha_2 \geq 0$ *such that* $\alpha_1 + \alpha_2 = 1$,

$$\alpha_1 x_1 + \alpha_2 x_2 = cP^T(\alpha_1 x_1 + \alpha_2 x_2) + (1-c)(\alpha_1 v_1 + \alpha_2 v_2) \tag{23}$$

Applying either Jeh or Haveliwala's method can help us pre-calculate the PPV of each CUI before the searching phrase. During the searching phrase, we only need to add up all the corresponding unit PPV to be the query's PPV. In this way, we can use the pre-calculated PPV to accelerate the search response.

7.3. Expected Impact

Effectively querying MEDLINE by PubMed is not an easy task for non-expert users. Our hybrid query expansion method for query the MEDLINE has greatly improved the recall-precision performance in biomedical information retrieval. However, our method is not limited to biomedical area. As long as there is suitable ontology graph, we

can apply our personalized PageRank query expansion into any area. In addition, we can apply the WEST algorithm if the hierarchy of ontology graph can be obtained.

Appendices

Appendix A: List of acronyms and abbreviations

AOD	-	Alcohol and Other Drug
API	-	Application Programming Interface
BIR	-	Biomedical Information Retrieval
CSP	-	CRISP Thesaurus
CUI	-	Concept Unique Identifier
DAG	-	Directed Acyclic Graph
GO	-	Gene Ontology
IC	-	Information Content
ICD10CM	-	Int'l Classification of Disease, 10^{th} edition, Clinical Modification
LCA	-	Least Common Ancestor
MAP	-	Mean Average Precision
MEDLINE	-	Medical Literature Analysis and Retrieval System Online
MC dataset	-	Miller and Charles dataset
MSH/MeSH	-	Medical Subject Headings
MTH	-	Metathesaurus MTH
NCBI	-	National Center for Biotechnology Information
NLM	-	National Library of Medicine
PPV	-	Personalized PageRank Vector
PRF	-	Pseudo Relevance Feedback
PS-IPF	-	PPV Score – Inverse PPV Frequency
RCD	-	Clinical Term Version 3
SNOMEDCT	-	SNOMED Clinical Term
SpecLev	-	Specification Level
TF-IDF	-	Term Frequency – Inverse Document Frequency
WEST	-	Weighted Edge Similarity Tools
WSD	-	Word Sense Disambiguation
UMLS	-	Unified Medical Language System

Appendix B: Public web services provided by WEST

B.1. Web Service

The following web services are provided and supported by WEST team at:

Uri: ‘urn:LiangSimilarity’

Proxy: ‘http://bioir.cs.clemson.edu:17581/’

B.2. Web Service Functions

```
double query(string word1, string word2[, string strategyCode[, float
alpha[, float beta]]])
```

Table 22: Strategy Code of WEST Web Service

Methods	Strategy Code
Weighted Edge Hybrid	wehybrid
Weighted Edge Sech	wesech
Weighted Edge tanhc	wetanhc
Li's Method	li

Example:

```
double res = query("boy", "man");  //Default Weighted Edge Hybrid with
alpha 0.85
double res = query("boy", "man", "wesech"); //Weighted Edge Sech
double res = query("boy", "man", "wesech", 0.87); //Weighted Edge Sech
with alpha 0.87
double res = query("boy", "man", "li", 0.2, 0.3); // Li's method with
alpha 0.2 and beta 0.3
```

B.3. Perl Client Sample using SOAP::Lite

```
use SOAP::Lite;
my $soap = SOAP::Lite
-> uri('urn:LiangSimilarity')
-> proxy('http://bioinformatics.clemson.edu:17581/');
```

```
my $res = $soap->query("boy", "man");
print "boy~man:".$res->result."\n";
```

B.4. PHP Client Sample using PHP::SOAP

```
$client = new SoapClient(NULL,
            array(
            "location" => "http://bioinformatics.clemson.edu:17581/",
            "uri"      => "urn:LiangSimilarity",
            "style"    => SOAP_RPC,
            "use"      => SOAP_ENCODED
          ));
$res = $client->query("boy", "man");
```

Appendix C: Install and Run BioIRWeb website

C.1 Installation

1. MetaMap

    - Need to install both MetaMap10 and MetaMap API (extract both into the /root/workspace/MetaMap)
    - Set environement PATH and JAVA_HOME in ~/.bashrc

      ```
      export JAVA_HOME=”/usr/lib/jvm/java-6-openjdk”
      export PATH=$PATH:/root/workspace/MetaMap/public_mm/bin
      ```

    - /root/workspace/MetaMap/bin/install

2. UKB_PPV:

. chmod of the UKB_PPV in the ukb_ppv directory

    a. install boost library

Install Tomcat

. the details follow the Tomcat and Eclipse document

C.2 Startup the web application

1. `cd /usr/share/tomcat6/bin`

   `sh shutdown.sh`

   In order to shut down the Tomcat run by the Ubuntu

2. Open Java Eclipse
3. Run the Eclipse's Tomcat server
    a. click BioIRWeb in the Eclipse's Package Explorer
    b. click the Green Triangle Button to run program with the right drop list
    c. select Run As->Run on Server
4. Open another terminal (startup the MetaMap daemon)

   ```
   cd ~/workspace/MetaMap/
   sh metamap_start.sh
   ```

   In the Web Browser, enter http://localhost:8080/BioIRWeb/index.jsp

   If any Java Null Pointer errors, check the Java Library *screenshot* in the BioIRWeb directory.

References

1. Jensen, L.J., J. Saric, and P. Bork, *Literature mining for the biologist: from information retrieval to biological discovery.* Nature reviews genetics, 2006. **7**(2): p. 119-129.

2. Islamaj Dogan, R., et al., *Understanding PubMed® user search behavior through log analysis.* Database (Oxford), 2009.

3. *Fact Sheet MEDLINE*. 2011; Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html.

4. *PubMed Website*. Available from: http://www.ncbi.nlm.nih.gov/pubmed/.

5. Hersh, W.R., *Information retrieval: a health and biomedical perspective*. 2009: Springer Verlag.

6. Bernstam, E. *MedlineQBE (Query-by-Example)*. 2001: American Medical Informatics Association.

7. McKibbon, K. and C. Walker Dilks, *The quality and impact of MEDLINE searches performed by end users.* Health libraries review, 1995. **12**(3): p. 191-200.

8. Wildemuth, B.M. and M.E. Moore, *End-user search behaviors and their relationship to search effectiveness.* Bulletin of the Medical Library Association, 1995. **83**(3): p. 294.

9. Hersh, W., et al. *OHSUMED: an interactive retrieval evaluation and new large test collection for research*. 1994: Springer-Verlag New York, Inc.

10. Hersh, W.R. and D. Hickam, *Information retrieval in medicine: the SAPHIRE experience.* Journal of the American Society for Information Science, 1995. **46**(10): p. 743-747.

11. Hersh, W.R. and D.H. Hickam, *A comparison of retrieval effectiveness for three methods of indexing medical literature.* The American journal of the medical sciences, 1992. **303**(5): p. 292.

12. Hersh, W.R., et al., *A performance and failure analysis of SAPHIRE with a MEDLINE test collection.* J Am Med Inform Assoc, 1994. **1**(1): p. 51-60.

13. Srinivasan, P., *Optimal document-indexing vocabulary for MEDLINE.* Information Processing & Management, 1996. **32**(5): p. 503-514.

14. Srinivasan, P., *Exploring query expansion strategies for MEDLINE.* Journal of the American Medical Information Association, 1995. **3**: p. 157-167.

15. Yoo, S. and J. Choi, *Improving MEDLINE document retrieval using automatic query expansion*, in *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*. 2007, Springer-Verlag: Hanoi, Vietnam. p. 241-249.

16. Abdou, S., P. Ruck, and J. Savoy, *Evaluation of stemming, query expansion and manual indexing approaches for the genomic task.* cell. **501**: p. 105.

17. Taylor, W.P. and J.Z. Wang. *Semantic Graph Based Document-Indexing Strategy for MEDLINE*. in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),*. 2010.

18. Dong, L., R.G. Smith, and B.G. Buchanan. *Automating the Selection of Stories for AI in the News*. in *proceedings of the Twenty Fourth International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, IEA-AIE*. 2011.

19. Dong, L., R.G. Smith, and B.G. Buchanan. *NewsFinder: Automating an Artificial Intelligence News Service*. in *23rd Annual Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI '11)*. 2011. San Francisco, CA.

20. Dong, L., P.K. Srimani, and J.Z. Wang, *WEST: Weighted-Edge Based Similarity Measurement Tools for Word Semantics*, in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. 2010, IEEE Computer Society. p. 216-223.

21. Dong, L., P.K. Srimani, and J.Z. Wang. *Weighted Edge: A New Method to Measure the Semantic Similarity of Words based on WordNet*. in *Global WordNet Conference* 2010.

22. Dong, L., P.K. Srimani, and J. Wang. *Ontology Graph based Query Expansion for Biomedical Information Retrieval*. in *IEEE International Conference on Bioinformatics and Biomedicine*. 2011. Atlanta, GA. USA.

23. Taylor, W.P. and others, *Creating a biomedical ontology indexed search engine to improve the semantic relevance of retreived medical text*, in *The Graduate School*. 2010, Clemson University.

24. *Ontology Wikipedia*. 2011; Available from: http://en.wikipedia.org/wiki/Ontology.

25. Spasic, I., et al., *Text mining and ontologies in biomedicine: making sense of raw text.* Briefings in Bioinformatics, 2005. **6**(3).

26. *Ontology Information Science*. 2011; Available from: http://en.wikipedia.org/wiki/Ontology_%28information_science%29.

27. Miller, G.A., *WordNet: a lexical database for English.* Communications of the ACM, 1995. **38**(11): p. 39-41.

28. Humphreys, B.L. and D.A.B. Lindberg. *Building the unified medical language system*. 1989.

29. Humphreys, B.L., et al., *The unified medical language system.* Journal of the American Medical Informatics Association, 1998. **5**(1).

30. Lipscomb, C.E., *Medical subject headings (MeSH).* Bulletin of the Medical Library Association, 2000. **88**(3).

31. Spackman, K.A., et al. *SNOMED RT: a reference terminology for health care*. 1997.

32. Stearns, M.Q., et al. *SNOMED clinical terms: overview of the development process and project status*. 2001.

33. *NCBI Entrez Taxonomy Homepage*. 2011; Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy.

34. *MeSH records*. Available from: http://www.nlm.nih.gov/mesh/intro_record_types.html.

35. *MeSH Tree Structure*. 2011; Available from: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/015_020.html.

36. Schuyler, P.L., et al., *The UMLS Metathesaurus: representing different views of biomedical concepts.* Bull Med Libr Assoc, 1993. **81**(2): p. 217-22.

37. Manning, C.D., P. Raghavan, and H. Schutze, *An introduction to information retrieval*. 2008: Cambridge University Press.

38. Miller, G.A. and W.G. Charles, *Contextual Correlates of Semantic Similarity.* Language and Cognitive Processes, 1991. **6**(1): p. 1-28.

39. *MEDLINE Wikipedia*. Available from: http://en.wikipedia.org/wiki/MEDLINE.

40. Krallinger, M. and A. Valencia, *Text-mining and information-retrieval services for molecular biology.* Genome Biology, 2005. **6**(7): p. 224.

41. *Query Expansion Wikipedia*. Available from: http://en.wikipedia.org/wiki/Query_expansion.

42. Gauch, S., J. Wang, and S.M. Rachakonda, *A corpus analysis approach for automatic query expansion and its extension to multiple databases.* ACM Transactions on Information Systems (TOIS), 1999. **17**(3): p. 250-269.

43. Billerbeck, B. and J. Zobel. *Techniques for efficient query expansion*. 2004: Springer.

44. Cui, H., et al., *Query expansion by mining user logs.* IEEE transactions on knowledge and data engineering, 2003: p. 829-839.

45. *Relevance Feedback Wikipedia*. Available from: http://en.wikipedia.org/wiki/Relevance_feedback.

46. Wu, Z. and M. Palmer, *Verbs semantics and lexical selection*, in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* 1994, Association for Computational Linguistics: Las Cruces, New Mexico.

47. Li, Y.H., Z.A. Bandar, and D. McLean, *An approach for measuring semantic similarity between words using multiple information sources.* Ieee Transactions on Knowledge and Data Engineering, 2003. **15**(4): p. 871-882.

48. Rada, R., et al., *Development and Application of a Metric on Semantic Nets.* Ieee Transactions on Systems Man and Cybernetics, 1989. **19**(1): p. 17-30.

49. Resnik, P., *Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.* Journal of Artificial Intelligence Research, 1999. **11**: p. 95-130.

50. Jiang, J.J. and D.W. Conrath. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy* in *Proc. ROCLING X*. 1997.

51. Lin, D. *An Information-Theoretic Definition of Similarity*. in *In Proceedings of the 15th International Conference on Machine Learning*. 1998: Morgan Kaufmann.

52. Resnik, P. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. in *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*. 1995.

53. Cilibrasi, R.L. and P.M.B. Vitanyi, *The Google similarity distance.* Ieee Transactions on Knowledge and Data Engineering, 2007. **19**(3): p. 370-383.

54. Gabrilovich, E. and S. Markovitch. *Computing semantic relatedness using Wikipedia-based explicit semantic analysis*. in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*. 2007.

55. Agirre, E. and A. Soroa. *Personalizing PageRank for Word Sense Disambiguation*. in *Proceedings of EACL-09*. 2009. Athens, Greece.

56. Miller, G.A., *Wordnet - a Lexical Database for English.* Communications of the Acm, 1995. **38**(11): p. 39-41.

57. Rubenste.H and Goodenou.Jb, *Contextual Correlates of Synonymy.* Communications of the Acm, 1965. **8**(10): p. 627-&.

58. Varelas, G., et al., *Semantic similarity methods in wordNet and their application to information retrieval on the web*, in *Proceedings of the 7th annual ACM international workshop on Web information and data management*. 2005, ACM: Bremen, Germany.

59. Pedersen, T., S. Patwardhan, and J. Michelizzi. *Wordnet::similarity - measuring the relatedness of concepts*. in *In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*. 2004. San Jose, CA.

60. Harel, D. and R.E. Tarjan, *Fast Algorithms for Finding Nearest Common Ancestors.* Siam Journal on Computing, 1984. **13**(2): p. 338-355.

61. Bender, M.A. and M. Farach-Colton, *The LCA problem revisited.* Latin 2000: Theoretical Informatics, 2000. **1776**: p. 88-94.

62. Bender, M.A., et al. *Finding least common ancestors in directed acyclic graphs*. in *12th Annual ACM-SIAM Symposium on Discrete Algorithms(SODA'01)*. 2001.

63. Steyvers, M. and J.B. Tenenbaum, *The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth.* Cognitive Science 29, 2005: p. 41-78.

64. Vladislav Daniel, V. and D.G. Wayne, *Mapping semantic relevancy of information displays*, in *CHI '07 extended abstracts on Human factors in computing systems*. 2007, ACM: San Jose, CA, USA.

65. McInnes, B.T., T. Pedersen, and S.V.S. Pakhomov. *UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity*. in *the Annual Symposium of the American Medical Informatics Association*. 2009. San Francisco, CA.

66. Leacock, C. and M. Chodorow, eds. *Combining local context with WordNet similarity for word sense identification*. WordNet:An electronic lexical database, ed. C. Fellbaum. 1998, MIT Press. 19-33.

67. Nguyen, H.A. and H. Al-Mubaid. *New ontology-based semantic similarity measure for the biomedical domain*. in *Granular Computing, 2006 IEEE International Conference on*. 2006.

68. *http://bioinformatics.clemson.edu/WEST.*

69. Matos, S., et al., *Concept-based query expansion for retrieving gene related publications from MEDLINE.* BMC bioinformatics, 2010. **11**: p. 212.

70. Agirre, E. and A. Soroa. *Personalizing pagerank for word sense disambiguation*. 2009.

71. Agirre, E., A. Soroa, and M. Stevenson, *Graph-based word sense disambiguation of biomedical documents.* Bioinformatics, 2010. **26**(22): p. 2889-96.

72. Agirre, E., et al., *A study on similarity and relatedness using distributional and WordNet-based approaches*, in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009, Association for Computational Linguistics: Boulder, Colorado. p. 19-27.

73. Agirre, E., et al. *Exploring Knowledge Bases for Similarity*. in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. 2010: European Language Resources Association (ELRA).

74. Ramage, D., A.N. Rafferty, and C.D. Manning. *Random walks for text semantic similarity*. 2009.

75. Mitra, M., A. Singhal, and C. Buckley. *Improving automatic query expansion*. 1998: ACM.

76. Brin, S. and L. Page, *The anatomy of a large-scale hypertextual Web search engine* 1.* Computer networks and ISDN systems, 1998. **30**(1-7): p. 107-117.

77. Page, L., et al., *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab., 1999.

78. Langville, A.N. and C.D. Meyer, *Deeper inside pagerank.* Internet Mathematics, 2004. **1**(3): p. 335-380.

79. Berkhin, P., *A survey on pagerank computing.* Internet Mathematics, 2005. **2**(1): p. 73-120.

80. Kamvar, S.D., et al., *Exploiting the block structure of the web for computing pagerank.* 2003.

81. Langville, A.N. and C.D. Meyer, *Fiddling with PageRank.* 2003.

82. Haveliwala, T., *Topic-Sensitive PageRank*. 2002, IEEE Transactions on Knowledge and Data Engineering.

83. Haveliwala, T., et al., *An Analytical Comparison of Approaches to Personalizing PageRank.* 2003.

84. Aronson, A.R. and F.M. Lang, *An overview of MetaMap: historical perspective and recent advances.* J Am Med Inform Assoc, 2010. **17**(3): p. 229-36.

85. Aronson, A.R., *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Proc AMIA Symp, 2001: p. 17-21.

86. Sparck-Jones, K., S. Walker, and S.E. Robertson, *A probabilistic model of information retrieval: development and comparative experiments Part 1.* Information Processing & Management, 2000. **36**(6): p. 779-808.

87. Voorhees, E.M., *Query expansion using lexical-semantic relations,* in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. 1994, Springer-Verlag New York, Inc.: Dublin, Ireland. p. 61-69.

88. Jalali, V. and M. Borujerdi, *Concept Based Pseudo Relevance Feedback in Biomedical Field Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, R. Lee and N. Ishii, Editors. 2009, Springer Berlin / Heidelberg. p. 69-79.

89. Jalali, V. and M.R.M. Borujerdi. *The effect of using domain specific ontologies in query expansion in medical field*. 2008: IEEE.

90. Li, Y., Z.A. Bandar, and D. McLean, *An approach for measuring semantic similarity between words using multiple information sources.* IEEE transactions on knowledge and data engineering, 2003: p. 871-882.

91. *UMLS Basics* Available from: http://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.htm.

92. Bos, L., *Medical and care compunetics 3*. 2006: IOS Press.

93. *Entrez*. Available from: http://www.ncbi.nlm.nih.gov/sites/gquery.

94. *MetaMapped Results Information*. 2011; Available from: http://skr.nlm.nih.gov/resource/MetaMappedBaselineInfo.shtml.

95. *2011 MetaMapped Medline Baseline Results*. Available from: http://mbr.nlm.nih.gov/Download/MetaMapped_Medline/2011/.

96. Jeh, G. and J. Widom. *Scaling personalized web search*. in *Proceedings of the Twelfth International World Wide Web Conference* 2003.