The Pennsylvania State University
The Graduate School
College of Information Sciences and Technology

EXPLORING APPLYING MESH ONTOLOGY FOR BIOMEDICAL PATENT CITATION RECOMMENDATION

A Thesis in
Information Sciences and Technology

by

Yinan Wan

© 2014 Yinan Wan

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2014
The thesis of Yinan Wan was reviewed and approved* by the following:

John Yen  
Director of Strategic Research Initiatives  
University Professor of Information Sciences and Technology  
Thesis Advisor

Zhen Lei  
Assistant Professor of Energy and Environment Economics

Dinghao Wu  
Assistant Professor of Information Sciences and Technology

Peter Forster  
Senior Lecturer, Security Risk and Analysis  
Assistant Dean for Online Programs and Professional Education, College of IST

*Signatures are on file in the Graduate School
ABSTRACT

Patent citation recommendation, as one of the prior-art search, is critical in the patent examination. Other than paper citation, the content scope is important for patent recommendation. A thorough prior-art search for patent will both help applicants to properly scale the proposal and USPTO examiners to evaluate the originality of the invention and accept how the invention can be distinguished from the prior art. Integrating technical terminology will help determine the scale of technical paper in a more precise way. MeSH ontology is a controlled vocabulary with hierarchical semantics developed by the National Library of Medicine to index the biomedical articles.

Here we developed a method to assign and evaluate the MeSH semantic similarity for the patent document. The experimental results generate a three-step “best measure set” to integrate the MeSH descriptor semantic similarity into the document similarity measure. We also further improved the result by using the Medical Text Indexer (MTI) to assign the MeSH descriptors.

We used 28438 Biotechnology patents as patent pool. Results evaluated by the average recall rate suggest though the MeSH semantic similarity measure may not exceed some of the sophisticated content-based measure, its outperforming the basic cosine similarity measure suggests its potential to act as one feature to be integrate into other complex search engine.
# TABLE OF CONTENTS

List of Figures .................................................................................................................. vi
List of Tables .................................................................................................................... vii
Acknowledgements ......................................................................................................... viii

Chapter 1 Introduction ...................................................................................................... 1

Chapter 2 Background and related work ......................................................................... 4
  Review of prior-art search ............................................................................................ 4
  Biomedical Ontology and MeSH .................................................................................. 6
  Applying Biomedical Ontologies in Information Retrieval ........................................ 9

Chapter 3 Problem Definition and Proposed Solution .................................................. 11
  Notation ......................................................................................................................... 11
  MeSH descriptor assignment ....................................................................................... 11
    Direct Extraction ....................................................................................................... 12
    Auto Assignment ....................................................................................................... 13
  Semantic similarity measure between two MeSH tree nodes .................................... 14
    Path finding measure ............................................................................................... 15
  Information content based measure .......................................................................... 16
  Semantics similarity measure between two MeSH descriptors .................................... 18
  Implementing Semantic Similarity in MeSH in patent recommendation .................... 19
    Applying semantic similarity in MeSH in patent recommendation ......................... 19
    Applying MeSH descriptor in content-based patent recommendation .................... 20

Chapter 4 Experiment and Result ................................................................................... 22
  Experimental set up ..................................................................................................... 22
    Data Set ..................................................................................................................... 22
    Evaluation Metrics .................................................................................................... 23
    Baseline methods ...................................................................................................... 24
  Experimental Result .................................................................................................... 24
    MeSH ontology semantic similarity measure analysis .............................................. 24
    Patent candidate selection analysis .......................................................................... 27
    MeSH descriptor Assignment Analysis .................................................................... 29
    MeSH descriptor in content-based similarity measure analysis ............................... 31
    Baseline comparison ............................................................................................... 33

Chapter 5 Discussion ...................................................................................................... 35
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline comparison</td>
<td>35</td>
</tr>
<tr>
<td>MeSH descriptor assignment</td>
<td>36</td>
</tr>
<tr>
<td>Information Loss</td>
<td>39</td>
</tr>
<tr>
<td>Chapter 6 Conclusion</td>
<td>40</td>
</tr>
<tr>
<td>Bibliography</td>
<td>42</td>
</tr>
<tr>
<td>Appendix</td>
<td>44</td>
</tr>
<tr>
<td>MeSH ontology</td>
<td>44</td>
</tr>
<tr>
<td>Implementing the semantic similarity measure</td>
<td>45</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2-1. Example of MeSH descriptor record: Stomach Neoplasms .................................. 8

Figure 2-2 Demonstration of positions where descriptor “Stomach Neoplasms” [D013274] located ........................................................................................................................................ 8

Figure 3-1 MTI Process Workflow[28] .......................................................................................... 13

Figure 4-1 Evaluation of MeSH ontology semantic similarity using different tree node and document similarity measure. ........................................................................................................... 25

Figure 4-2 Performance of applying Wu&Palmer as step-1 method in patent citation recommendation. ........................................................................................................................................... 28

Figure 4-3 Distribution of number of MeSH descriptors assigned to each patent. ................. 29

Figure 4-4 Performance of using MeSH descriptors from Direct extraction and MTI auto assignment. ........................................................................................................................................... 31

Figure 4-5 Evaluation of applying MeSH descriptors and top-k keywords in content-based similarity measure. ........................................................................................................................................... 32

Figure 4-6 Performance evaluation: MeSH semantics and baseline comparison ..................... 33

Figure 5-1 Example of MeSH terms in a PubMed record with PMID: 20373513 ................. 37
LIST OF TABLES

Table 4-1 Precision of eight semantic similarity measure with selected k-value.................... 25
Table 4-2 Precision of eight semantic similarity measure with selected k-value.................... 26
Table 4-3 The average recall rate of top-k candidates for the combination queries............... 32
Table 5-1 Ranking of “true references” for the testing patent “7217537”......................... 35
ACKNOWLEDGEMENTS

I would like to thank Dr. Yen for accepting me as a Master student under his supervision, generously offer advice and support both for the thesis and the two-year study. Thanks to Dr. Zhen Lei for his advices and suggestions during all the past year on this project. I would also like to thank Dr. Prasenjit Mitra and Dr. Wang-Chien Lee for the critical analysis and recommendation during the early and later discussion of the project. Special thanks go to Sooyoung Oh for all the help with the managing the dataset and operation system, as well as all the suggestions offered as a peer researcher. Finally I feel grateful to Dr. John Yen, Dr. Zhen Lei and Dr. Dinghao Wu for serving as my committee members.
Chapter 1

Introduction

The rapid growing in the information technology development is impacting Intellectual Property field. Patent citation recommendation, one of the prior-art search, is critical in the patent application both for the applicants and United States Patent and Trademark Office (USPTO) examiners. A thorough prior-art search will help the patent applicant to narrow down the candidate patent into a proper scope that are both not too wide to be patented and innovated enough to be applied as a new patent. On the other hand, a precise and comprehensive prior-art search will help the USPTO examiners to decide whether the application is patentable and scale down the application into proper scale with supplementary references.

Patent documents are viewed as a kind of technical document, thus, integrating the technical terminology may help describe the patent more precisely and improve the recommendation. MeSH ontology is a controlled vocabulary with hierarchical semantics developed by the National Library of Medicine (NLM) to index the articles in the MEDLINE database. In the MEDLINE database, MeSH descriptors are pre-assigned to each article by the indexing experts. Usually each article is assigned 5–15 descriptors. The MeSH ontology is originally developed for article indexing and cataloging only in the PubMed, which is the major portal in NLM for the search of a large number of biomedical articles including MEDLINE database. However, most of the studies have focused on applying the MeSH semantics into MEDLINE articles retrieval. But few studies have been found to applying the MeSH ontology outside MEDLINE.

In this work we propose to apply the MeSH ontology into the patent citation recommendation. Since one MeSH descriptor usually contains more than one MeSH tree node in
the hierarchical structure, and one document is most likely to be assigned with several MeSH
descriptors, we proposed a three-step semantic similarity measure.

First we compared four methods of computing semantic similarity between two tree
nodes; for the second step, we proposed to use the Average Maximum Measure (AMM) to
measure the similarity between two MeSH descriptors; and the final step, we propose both the
AMM and a modified tf-idf related AMM measure. The MeSH descriptors are originally
assigned by directly extracted from the claim of each patent document. To further improve the
similarity measure, we also apply the Medical Text Indexer (MTI) for MeSH descriptor auto
assignment.

We apply this similarity measure into 28438 Biotechnology patents, and randomly
selected 100 patents from this patent pool as the testing set. For evaluation purpose, all the
patents in the testing set have at least four reference patents. Since in this task, it is more
important to not miss any patents, the patent citation recommendation can be viewed as recall-
oriented. We use the average recall rate as the evaluation matrix. Average precision rate is also
listed for reference.

The rest of the thesis is organized as below:

Chapter 2 will provide the necessary background and related work both about the citation
recommendation and the existing application of MeSH ontology into MEDLINE and other
databases.

Chapter 3 will describe and specify the MeSH semantic similarity in patent documents,
introduce our three-step similarity measure and our candidate methods in each step. This chapter
will also describe the evaluation matrices and two baseline methods for comparison purpose.

Chapter 4 will describe the dataset for the experiment and compare the results using
different methods.
Chapter 5 will describe the differences in results from the experiments of Chapter 4. It also will draw the underline issues related with each method.

Chapter 6 will highlight the conclusion from the study and propose the future work.

Appendix will describe some technical issues and problems in details during the implementation of the method and evaluation.
Chapter 2

Background and related work

This text is in the type style called “Normal” and should be used for the body of your thesis. Great as the differences are between the breeds of pigeons, I am fully convinced that the common opinion of naturalists is correct, namely, that all have descended from the rock-pigeon (Columba livia), including under this term several geographical races or sub-species, which differ from each other in the most trifling respects.

Review of prior-art search

Since the patent is playing a key role of the Intellectual property protection, with the growing development in technologies, a remarkable number of patent applications, including patent retrieval, have been developed in the past decade. Prior-art search, as one of the main types in the patent retrieval, has become increasingly critical yet a challenging task to the United States Patent Trademark Office (USPTO) patent examiners.

Prior art, also known as state of the art, in most system of patent law, includes all the information that are relevant to the patent, or to be more specifically, the originality of the patent’s claim. The information should have already been declared as to be public before a given date [1]. A prior art search involves search in multiple databases to obtain knowledge that relates to the patent, including previous patents, trade journal articles, publications and so on. An inventor has a legal obligation to disclose any relevant information they acquire about their inventions to the USPTO. A patent search can be viewed as a subset of prior-art search which researchers focused specifically on the relevant patents. Here we refer prior-art search specific to
the area of patent search. A good prior-art search can facilitate the inventors to validate their ideas and decide whether it is patentable, and narrow down the proposal to a suitable scale that is innovative enough to apply as a new patent. When applying for a new patent, the submission of relevant prior art will help the USPTO to evaluate the originality of the invention and accept how the invention can be distinguished from the prior art. A thorough study and submission of prior art work will also protect the inventors from a lot of challenges about the novelty and validity as many parties who try to infringe a patent by providing prior-art that is not included in the patent.

For USPTO patent examiners, on the other hand, the search of prior-art will help them to decide whether this invention is novel and can be patented, whether there’s any prior-art work that is closely related to the invention but not listed in the reference field, or if the claims should be modified to a more proper scope. Those relevant patents, from both the submission of the inventors and added by the USPTO examiners, will be advised to be listed under the Reference field. Multiple facts, including the limited time and resources, may lead the examiners to miss some important prior-art patents, which may further lead to the fact that the to-be patented inventions might miss some important references, cite improper patents, or the claim may be too broad to suit the invention.

All the above issues have drawn the public attentions on developing innovative prior-art search methods in order to improve the current situation. The existing work about prior-art search focuses on two major domains: information retrieval and citation network.

Patent recommendation work based on information retrieval mostly focuses on keyword search, where mostly the retrieval models will extract keywords from the patents to form the queries as the representation of the whole document. Thus the success of retrieval largely depends on the proper choice of the queries. However, writers of the patent documents tend to intentionally use many vague terms and expressions to cover as most of the invention scope as possible. [2] Hence, generating proper queries or representations as the expansions becomes a
critical task. Larkey[3] introduced patent classification for retrieval and first developed a system for USPTO. Later Takaki[4] et al. studied the claim part of the patent document, discovering multiple subtopics and assigning one sub-query for each. The final retrieval is done by adding relevance score from each sub-queries together with weight assigned according to the importance of the subtopic. Study from Xue and Croft [5] however, demonstrates queries generating from summary field might work better than the others. More recently, Bash and Rauber [2] proposed an approach to increase the irretrievability of patents by expanding the queries with pseudo relevance feedback. By checking the similarity with query patents, patents for relevance feedback are identified for further reference of missing terms in the query patents.

**Biomedical Ontology and MeSH**

While most of the work studying the query generation for patent prior-art search have been focused on the pure content, while there are other important external resources related with technical papers.

Ontologies can be viewed as general tools of information to represent a subject. Domain specific ontologies further can be regarded as a set of technical terms characterizing the specific domain.[6] The most common biomedical domain ontologies include UMLS and MeSH.

The Unified Medical Language System, known as UMLS, is a set of files and software that brings together many health and biomedical vocabularies and standards to provide a unified yet comprehensive definition and information interpretation in the healthcare field.[7] Specifically, the UMLS Metathesaurus (refer to as UMLS in the paper) is a very large, multi-purpose vocabulary databases containing information about the biomedical and health related concepts, their alternative name and relationship between them. The source vocabularies of the UMLS come from multiple thesauri including lists of controlled terms from various healthcare
and clinical fields such as diseases definition, medical devices and expert diagnostic systems. The concepts from over 168 biomedical vocabularies including MeSH define the scope of UMLS. The UMLS is organized by concept. To identify the different names for the same concepts and connect multiple vocabularies, the system assigns unique, permanent identifiers to concepts (CUI) and concept names it contains. [8] To be noticed, since the scope of each individual vocabulary in the UMLS might be various, one record in a particular vocabulary could contain multiple UMLS concept due to the fact that the record may be defined to be broader than all the UMLS concepts it contains.

Medical Subject Headings, or MeSH is one of the ontologies in UMLS. MeSH is a controlled vocabulary thesaurus defined by the National Library of Medicine (NLM). [9]. MeSH ontology is originally designed to annotate and index the articles in the MEDLINE database [10], a corpus made up of abstracts of biomedical journal articles, but has been widely used all kinds of NLM-produced database of books, documents and general library indexing. It has also been used in other areas like information retrieval for the medical record and expert recommendations [11][12].

Most records in the MeSH is called a “descriptor”, which has a unique descriptor ID followed by a descriptor name (also known as MeSH heading). Normally each descriptor will have several synonyms, and MeSH uses a three-level structure: Descriptor/Concept/Term to describe the synonymous relationship. Figure 2.1 is a typical record of a descriptor, as it’s shown in the figure, the descriptor “Stomach Neoplasms” have three concepts, the first one is the preferred concept which have the same name as the descriptor, while the other two concepts can be new as synonyms, and the scopes are either broader or narrower comparing with the descriptor. Similarly one concept will have several terms (also called entry terms), all synonymous to each other with one of them to be the preferred term having the same name as the concept. In general, all the terms in one descriptor can be viewed as the synonyms of the descriptor. To be noticed,
the “entry term” in the MeSH record is a UMLS term, and each of them has a Concept Unique ID (CUI). So in figure 2.1, the descriptor “Stomach Neoplasms” will have 9 UMLS terms.

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Stomach Neoplasms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Number</td>
<td>D01, 234, 478, 767</td>
</tr>
<tr>
<td>Tree Number</td>
<td>C01, 301, 371, 567</td>
</tr>
<tr>
<td>Tree Number</td>
<td>C06, 405, 249, 367</td>
</tr>
<tr>
<td>Tree Number</td>
<td>C06, 405, 748, 389</td>
</tr>
</tbody>
</table>

**Annotation**

- coord IM with histol type of nopo (IM); nopo of ruminant stomach or its part coord STOMACH; RUMINANT or indention (IM) with STOMACH NEOPLASMS / bone (IM) + histol type of nopo /cui (IM)

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Stomach Neoplasms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope Note</td>
<td>Tumors or cancer of the STOMACH.</td>
</tr>
<tr>
<td>Term</td>
<td>Stomach Neoplasms</td>
</tr>
<tr>
<td>Term</td>
<td>Gastric Neoplasms</td>
</tr>
<tr>
<td>Term</td>
<td>Neoplasms, Gastric</td>
</tr>
<tr>
<td>Term</td>
<td>Neoplasms, Stomach</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 2</th>
<th>Cancer of Stomach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>Cancer of Stomach</td>
</tr>
<tr>
<td>Term</td>
<td>Cancer of the Stomach</td>
</tr>
<tr>
<td>Term</td>
<td>Gastric Cancer</td>
</tr>
<tr>
<td>Term</td>
<td>Stomach Cancer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 3</th>
<th>Gastric Cancer, Familial Diffuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>Gastric Cancer, Familial Diffuse</td>
</tr>
</tbody>
</table>

**Allowable Qualifiers**


**Entry Version**

- STOMACH NEOPL

**Date of Entry**

- 19990101

**Unique ID**

- D013274

Figure 2-1. Example of MeSH descriptor record: Stomach Neoplasms

![MeSH Tree Diagram](image)

Figure 2-2 Demonstration of positions where descriptor “Stomach Neoplasms” [D013274] located

There are 27,149 descriptors in the 2013 MeSH, which are organized into 16 categories: category A for anatomic terms, category B for organisms, C for diseases, D for drugs and
Each category is an individual hierarchical tree structure, where more general terms appear at the top and more specific ones stay at the bottom. Each position (or node) in the tree is coded with a combination of alphabets and numbers. For example, the node C04.588.274.476.767 refers to a node under the disease category at level five. The numbers serve only to locate the node in each tree, but having no intrinsic significance. Specifically, the same number at different level or different categories have to relationship unless otherwise indicated. Also the size of two numbers does not have significant means, actually at each level the node is indexed alphabetically. To be noticed, since most time one MeSH descriptor contains more than one UMLS terms, leading to the fact they could be categorized into multiple positions in the trees. Thus, each MeSH descriptor will have at least one tree number, but more likely to have more. In figure 2.1, the descriptor “Stomach Neoplasms” have four tree numbers: C04.588.274.476.767, C06.301.371.767, C06.405.249.767 and C06.405.748.789, which indicate this descriptor appears four times in the MeSH structure, specifically all under the category of diseases (C). Figure 2-2 demonstrates how the four tree nodes of descriptor “Stomach Neoplams” located in the MeSH structure.

Applying Biomedical Ontologies in Information Retrieval

MeSH ontology is originally designed for indexing the journals in the MEDLINE database. PubMed, which is the online search engine for biomedical journal and citations from MEDLINE and several other major databases, use MeSH terms as one of the search fields in the query. Although MeSH terms in the PubMed is only used for cataloging and MeSH terms are queried with exact match, a lot other applications for biomedical article retrieval in MEDLINE integrating the MeSH semantics has been developed. Abasolo and Gomez [13] developed an MELISA, an ontology-based biomedical information retrieval prototype MELISA to expand the
PubMed search by applying the MeSH semantics to the MeSH field of the query. Zhu et.al [22] combined both MeSH semantic similarity and content-based similarity to generate the similarity matrix and improve the MEDLINE documents clustering. Lee et.al [12] combined the MeSH terms with the author information for expert and collaboration recommendation. Most of the researches are based on MEDLINE database, whose documents are manually pre-assigned with proper MeSH terms. Ruiz and Neveol [15] incorporated the MeSH terms for Medical Image retrieval, they used Medical Text Indexer (MTI) to index the annotation of each image with MeSH descriptors, and combining the image retrieval model they developed to further improve the result. This model treat MeSH descriptors as simple word vectors and used cosine normalization for the similarity calculation. Taschwer [16] applied the MeSH semantics for Medical case retrieval by extracting and expanding the MeSH descriptors to the patient record, combining with content-based measure. Although this strategy only slightly improve the retrieval performance, this research indicates the possibility of assigning and applying the MeSH ontology to full text retrieval.

Although the MeSH ontology has been developed and applied to the MEDLINE database for more than 40 years, most of the research and development are still focused on MEDLINE database, with some studies expanded to short text retrieval, like medical image or gene annotations. It is still challenge to apply the MeSH ontology to full-text retrieval like patent document.
Chapter 3

Problem Definition and Proposed Solution

As stated, the ultimate goal is to apply the MeSH ontology into the patent recommendation. In general, one patent document will be assigned several MeSH descriptors, while one MeSH descriptor usually contains more than one tree node. As a result, computing the MeSH semantic similarity between two patent documents will contain three steps: 1> semantic similarity measure between two MeSH tree nodes; 2> semantic similarity computation between two MeSH descriptors and 3> MeSH semantic similarity computation between two patent documents.

Notation

In this work, we consider each patent contains a set of MeSH terms, let \(D = \{M_1, M_2, \ldots, M_n\}\), where \(M_i\) represents one MeSH term in a document \(D\); and similarity between two document is presented as \(\text{sim}(D_1, D_2)\); while each MeSH term contains a set of nodes, let \(M = \{c_1, c_2, \ldots, c_n\}\), where \(c_i\) represents one tree number (node) of a MeSH term \(M\); similarity between two MeSH term is presented as \(\text{sim}(M_1, M_2)\), and in the end, similarity between two nodes are represented as \(\text{sim}(c_1, c_2)\).

MeSH descriptor assignment

MeSH ontology is first applied in the MEDLINE database, where MeSH descriptors are manually or semi-manually assigned to each article in the MEDLINE database. However here it
is impossible to set up a similar complex rules to employ manual MeSH descriptor assignment. As a result, we apply below two different ways to auto assign MeSH descriptors into each patent document.

**Direct Extraction**

The most direct way to assign MeSH descriptor to each patent document is to extract from the document. As described above, each MeSH descriptor may have multiple entry terms. Some of the entry terms can be considered as synonyms of the descriptors, while other terms are simply the same phrase of the descriptor but in a different morphological form. For example, the descriptor Abattoirs [D000003] also contains following entry terms: abattoirs; abattoir; slaughterhouses; slaughterhouse, where the first two are simply in different morphological forms of the main descriptor, while the later two are completely different words (and they are morphological forms within themselves). In this way, stemming is even not necessary, all the extracted MeSH descriptors are exact match of either the main descriptors or their entry terms.

Also, only 24% (26356 out of 108098 descriptors) of the MeSH descriptors are single word, all the others are phrases made up of multiple words. Considering the fact that descriptors made up of more than four words, for example, Gram-Negative Anaerobic Straight, Curved, and Helical Rods [D016965], are more likely to be short sentences describe a situation rather than phrases. Here to reduce the consuming of time computational resource we only count the descriptors with no more than four words into our MeSH descriptor pool. For descriptors having both no-more-than-five words and above-five words, like persistent fetal circulation syndrome [D010547], which has both entry term “alveolar capillary dysplasia with misalignment of pulmonary veins and other congenital anomalies” of 12 words and entry term “ACDMPV” of only one word, we keep the descriptor record, and all the terms with no more than four words, but
throw away terms of longer phrases. As a result, we keep 99149 records, which is 91% of the original MeSH vocabulary.

Auto Assignment

We also proposed another method of assigning MeSH descriptors to the patent document, by using NLM developed tool Medical Text Indexer (MTI)[28]. MTI is developed to recommend MeSH descriptors to assist indexers making the final decision. NLM even expanded MTI’s role to be the first-line indexer (MTIFL) for a few journals since 2011, which means it will independently assign the MeSH descriptors to the articles without further intervention by human indexer. In this way, it is reasonable to trust the result of auto assignment performed by MTI.

**Figure 3-1** MTI Process Workflow[28]

MTI perform the MeSH descriptor assignment by three steps: candidate retrieval, MeSH ontology restriction and list reorder. MetaMap [29] is the core algorithm for MTI process.
MetaMap generalize the UMLS terms from text and ranking by the frequency and relevance. Then the UMLS terms are restricted to MeSH descriptors only according to the MeSH ontology to provide candidates. Clustering and Reranking is applied, with extra filtering to provide the final list. At the candidate recommendation step, an alternative way will be extracting MeSH descriptors from the documents from the same algorithm of the PubMed Recommendation System. But candidates from this method are assigned with less weight as this is considered to be less confident.

**Semantic similarity measure between two MeSH tree nodes**

Semantic similarity uses a quantified score to estimate the similarity between two concepts. Measures of semantic similarity have proven useful in a number of Natural Language Processing tasks. With domain specific ontologies and resources in biomedical field, it is possible to adapt the measurement into biomedical texts for knowledge discovery and hypothesis.

Technically semantic similarity and relatedness are two different concepts. Specifically, semantic similarity is based on the *is-a* relation, or taxonomic relatedness; while semantic relatedness is a more general notion which can also refer to other relation like *part-of, treated-by* etc. In other words, semantic similarity can be viewed as a special case of semantic relatedness.[17-18] However, most time the terms similarity and relatedness are used interchangeably especially in biomedical domain, partially due to the fact a lot of the measures are applied to the same types tasks, for example measuring the distances between words using ontologies or other hierarchy.

As stated before, most of the MeSH descriptors will possess more than one tree nodes in the MeSH hierarchical structure. Here in section 3.2, I first present and discuss some existing
measures to evaluate the similarity between two tree nodes. Similarity between two MeSH descriptors and two patent documents will be further defined in the later sections.

**Path finding measure**

When concepts are organized in hierarchy such as the MeSH tree structure presented here, the concepts that are more near the roots of the tree, more general it tends to be; on the other hand, the concepts near the leaves are more specific. In this way, the most direct way to measure the similarity between two nodes in one tree is according to the path lengths.

Rada et al [19] first developed a “semantic distance measure” based on path lengths between concepts in the MeSH ontology. They simply rely on the “broader than ” relationship, travelling from one concept to the other by linking to the successors or ancestors. Mathematically this similarity measure between concepts can be defined as following:

\[
\text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{\text{path}(c_1, c_2)}
\]

where \(\text{path}(c_1, c_2)\) is the shortest path between \(c_1\) and \(c_2\). The path is considered to be the number of nodes between two concept in the MeSH hierarchical structure.

Wu and Palmer [20] present a measure of similarity based on the most specific concept which is the subsumer of both concepts to be measured. This is also called their least common ancestor:

\[
\text{sim}_{\text{up}}(c_1, c_2) = \frac{2 \times \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}
\]

where \(\text{lcs}(c_1, c_2)\) represents the least common ancestor of \(c_1\) and \(c_2\); \(\text{depth}(c)\) is the path between node \(c\) and root.

Leacock and Chodorow [21] further developed the a normalized path length measure based on the ratio of distance between two nodes to the size of the entire MeSH tree
\[
sim_{lc}(c_1, c_2) = \frac{-\log(\frac{\text{path}(c_1, c_2) + 1}{2 \times \max_{c \in \text{MeSH}} \text{depth}(c) + 1})}{\log(25)}
\]

Specifically, MeSH tree goes further to twelve levels at most. So in practice, \(\max_{c \in \text{MeSH}} \text{depth}(c) = 12\); to avoid \(\log(0)\) (when \(c_1 = c_2\)); we add 1 to both numerator and denominator; further, to normalize the score between 0 and 1, we divide the score with \(-\log(1/25)\) [22]. In this way the LC measure can be further written like:

\[
sim_{lc}(c_1, c_2) = \frac{-\log((\text{path}(c_1, c_2) + 1)/(2 \times 12 + 1))}{\log(25)} = 1 - \frac{\log(\text{path}(c_1, c_2) + 1)}{\log(25)}
\]

\[(3.4)\]

**Information content based measure**

The limitation of path finding based measures is that the degree of each node cannot be implied from the measure. Specifically, two general concepts can be connected with much fewer links than two specific concepts, while more specific concepts can deliver much more information.

Information content (IC) is a measure of specificity of a concept. Typically higher IC score is associated with more specific concepts; while more general concepts is more likely to have lower IC-score.

Corpus-IC[22] is the most common measure, based on the frequency counts of concepts found in a corpus of the text:

\[
IC_{\text{corpus}}(c) = -\log\left(\frac{\text{freq}(c)}{\text{freq}(\text{root})}\right)
\]

\[(3.5)\]
where \( \text{freq}(c) \) is the represents the frequency of the concept \( c \) in the corpus. It can be further modified using other statistics which can represent the weight or importance of the concept, like TF-IDF[23].

Resnik[24] first proposed applying the concept information content measure (IC measure) in semantic similarity:

\[
sim_{re}(c_1, c_2) = IC(lcs(c_1, c_2))
\] (3.6)

Here the similarity is computed by measuring the Information content of the least common ancestor between two concepts.

Lin [25] further develop a IC based measure for semantic similarity, scaling the information content of the least common ancestor by the information content of each individual concept, which is modified from the Wu&Parmer path finding measure [20]:

\[
sim_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}
\] (3.7)

Jiang and Conrath [26] defines the minimum distance between two concepts as:

\[
dist_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2))
\] (3.8)

Sanchez and Batet further modified the Leacock and Chodorow[6] path finding measure in equation (3.3) adapting into the IC measure, redefining the maximum depth as \( \text{ic}_{max} \), which according to the intrinsic information content definition, should be the top node which has the most leaves (D12); and the path between two concepts as Jiang and Conrath distance in equation (3.8). In this way, Sanchez and Batet’s measure can be described as:

\[
sim_{JC}(c_1, c_2) = e^{\frac{dist_{JC}(c_1, c_2)}{\lambda}},
\] (3.9)

All the measures mentioned above are scaled into scope between 0 and 1, where the higher score means to be more similar. And even though MeSH ontology have 16 subcategories, which each one is considered to be on individual hierarchical tree, we treat the whole MeSH
ontology to be one big tree by adding one general root above all the subcategories. This is because some MeSH descriptors have multiple tree numbers, which reside under different categories, if treating each subcategories separately, one will retrieve multiple similarity score under different scope. Also for two nodes residing under different subcategories, we will still able to retrieve their least common ancestor (LCS) if the general root is presented (which the LCS will be the general root). In this way, we will manually add one general root above all the subcategories.

**Semantics similarity measure between two MeSH descriptors**

As there are a lot of mature semantic similarity measures between two MeSH tree nodes, the measure between two MeSH descriptors are yet to be discussed.

Wang et al. [27] developed an Average Maximum Match (AMM) to measure the semantic similarity between two genes, annotated by two sets of GO (Gene Ontology [http://www.geneontology.org/](http://www.geneontology.org/)) terms. This is the similar situation as MeSH, which is comparing two MeSH descriptors described by a set of MeSH tree nodes. This measure has been reported to be consistent with human conception. Adapted to the MeSH, the AMM measure will be presented as following:

\[
\text{sim}(M_1, M_2) = \frac{\sum_{c_1 \in M_1} \max_{c_2 \in M_2} (c_1, c_2) + \sum_{c_1 \in M_2} \max_{c_2 \in M_1} (c_1 + c_2)}{|M_1| + |M_2|}
\] (3.10)

Where \( M \) is a MeSH descriptor represented by a set of tree nodes \( c, M = \{c_1, c_2, ..., c_m\} \); and \(|M|\) is the number of nodes in the set \( M \). AMM first defines the similarity between one MeSH tree node \( c_{2i} \) and a MeSH descriptor \( M_1 \) by choosing the score of tree node \( c_{2i} \) to the node in the \( M_1 \) that have the highest similarity, then sum up all the tree node similarity score and normalize by the
number of nodes in the descriptor set. This MeSH descriptor measure is first adopted by Zhu et al[22].

Implementing Semantic Similarity in MeSH in patent recommendation

Applying semantic similarity in MeSH in patent recommendation

Similarly, we consider a patent document to be a set of descriptors, so the above semantic similarity measure between two MeSH descriptors (Equation xx) can be further adapted to the documents:

$$
sim(d_1, d_2) = \frac{\sum_{M_1 \in d_1} \max_{M_2 \in d_2}(M_1, M_2) + \sum_{M_2 \in d_2} \max_{M_1 \in d_1}(M_1, M_2)}{|d_1| + |d_2|}
$$

(3.11)

Where M a MeSH descriptor, and d is a patent document represented by a set of MeSH descriptors, d=\{M_1, M_2, ..., M_m\}.

However, unlike MeSH descriptors where the tree nodes it contains have no difference to the MeSH descriptor according to the MeSH ontology, the MeSH terms in a patent document can be weighted in multiple ways. The most commonly adapted weight measure is TF-IDF [23].

MeSH terms are different from the keywords in the way that as technical term, they may not appear as frequent as other terms, but serve as critical terms in technical papers since they represent the domain specific information. However, within the terminology, each term should still be treated differently as most of them are still extracted or leveraged from the content. As a result, TF-IDF weight can still be integrated into the semantic similarity measure when using MeSH terms directly extracted from the document:

$$
sim(d_1, d_2) = \frac{\sum_{M_1 \in d_1} \alpha_1 \times \max_{M_2 \in d_2}(M_1, M_2) + \sum_{M_2 \in d_2} \alpha_2 \times \max_{M_1 \in d_1}(M_1, M_2)}{|d_1| + |d_2|}
$$

(3.12)
Where $\alpha$ is the tf-idf of the MeSH descriptor that one tree node is comparing to. It is computed as the following:

$$\alpha = tf \times idf = (0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d): w \in d\}}) \times \log \frac{|D|}{|d \in D : t \in d|}$$  \hspace{1cm} (3.13)

Where $f(w,d)$ is the frequency of any word $w$ in the document $t$; specifically $f(t,d)$ is the frequency of the MeSH descriptor $t$, summing up the frequency of all the entry terms in $t$ that appear in the document $d$.

In conclusion, the semantic similarity measure between two patent documents is a three-step calculation: 1> Calculating the semantic similarity between two tree nodes 2> Calculating the similarity between two MeSH descriptors based on two sets of tree nodes belonging to each MeSH descriptor and 3> Calculating the similarity between two patent documents based on two sets of MeSH descriptors assigned to each patent. The final result to represent the similarity between two patents is a related similarity score, the higher the score is, more similar two patents are at this scale.

**Applying MeSH descriptor in content-based patent recommendation**

An alternative approach to evaluate the value of MeSH ontology in patent recommendation is to combine and integrate the MeSH descriptors into content-based information retrieval method.

Indri [30] is a text-based information retrieval model combining the language model and inference network. The language model assigns a probability to a sequence of words by means of a probability distribution. Retrieved model from Indri are ranked based on probability the language model would generate from the terms of query, and this will be interpreted as the
likelihood of a document being relevant given a query. Indri allows complex queries including evidence combination and has the ability of specifying a wide range of constraints based on the complexity of document structure and customer’s design.

Top-k keywords ranked by tf-idf is a common query used for search engine. For top-k keywords, each word is assigned a weight using the tf-idf:

\[ Q_{cnt} = \beta \times W \]  

(3.14)

where \( W \) is the top-k keywords and \( \beta \) is tf-idf of each word. When applying to MeSH descriptors, similar as we compute tf-idf for MeSH descriptors above for semantic similarity measure, all the entry terms (UMLS terms) are aggregated together as one MeSH record:

\[ Q_{MeSH} = \nu \times \text{syn}(M) \]  

(3.15)

where \( \text{syn}(M) \) is a set of all the entry terms in a MeSH record \( M \), so during retrieval, the search engine will consider all the entry terms in \( M \) to be the same as the main descriptor term; \( \nu \) is the tf-idf of the MeSH descriptor, which using all the entry term frequency summing up as the frequency of the main descriptor.

We used the a second query here combining both MeSH descriptor and top-30 keywords:

\[ Q_{\text{combine}} = \alpha \times (\beta \times W) + (1- \alpha) \times \max(\beta) \times \text{syn}(M) \]  

(3.16)

where first part is the same as \( Q_{cnt} \), \( W \) is the top-30 keywords, \( \beta \) is the tf-idf for each keywords; for the second part, \( \text{syn}(M) \) is still the aggregating strategy that treating all the entry terms as the main MeSH, while different from \( Q_{MeSH} \), we are giving all the MeSH descriptors same weight, the highest tf-idf in top-30 keywords, which suggests we weigh the MeSH term same as the highest tf-idf keywords. The \( \alpha \) is the user-controlled parameter, suggesting how much we are depending on top-30 keywords and MeSH descriptors.
Chapter 4

Experiment and Result

We evaluated the solutions proposed above by applying the semantic similarity measure to a subset of U.S Biotechnology patents.

Experimental set up

Data Set

The dataset used to evaluate MeSH semantic similarity measure is a subset from the real dataset of U.S patents ranging from 1977 to 2010. The category of “Drug&Medical” based on the Patent Classification System [31] is selected for experiment. In general there are four subcategories within the “Drug&Medical” category, Drugs, Surgery&Medical Instruments, Biotechnology and Miscellaneous-Drug&Med. Since MeSH ontology is primarily built for journal articles in PubMed database aiming at biomedical research, the two sub-categories Biotechnology, which contains 28438 patents and Drug, which contains 28438 patents are selected as the patent pool in the experiment. Specifically the 28438 patents in Biotechnology set are treated as core patent pool, and the larger dataset of 87193 patents are labeled as general patent pool.

When testing the method within one particular patent pool, 100 patents are randomly generated from the patent pool to be the query set. The references in each patent are treated as the “ground truth”. Because of this setup, the extra criterion for the query set is to have at least four citations for the evaluation purpose.
MeSH terms are assigned using two sets of elements in the patent documents. One way is to assign the MeSH term based on only claims, which has been commonly believed to be the most text candidate for query generation. [2][4] However, since the MTI [24] is developed to generate MeSH descriptors from the abstract part of the articles, also more and more research indicates summary may provide more precise information for query generation, and further considering the particularity of MeSH as technical ontology, a second way used to assign the MeSH descriptor is to use content from title, abstract, claim and description adding up together.

**Evaluation Metrics**

For each patent in the testing set, a similarity score is computed for each document in the patent pool. The semantic similarity is computed by using the three-step computation and a final score will be retrieved. Then the documents are listed in the descending order and the top k patents are retrieved as the candidates. The average precision Pres@k (3.17) and recall Rec@k (3.18) are computed for the query set. Since patent citation recommendation is a recall-oriented task as it is more important to not miss any references than retrieve more relevant documents, more attention should be focused on the recall performance.

\[
\text{Pres@k} = \frac{\sum \frac{\# \text{hit} - \text{in} - \text{topK}}{k}}{N} \quad (3.17)
\]

\[
\text{Rec@k} = \frac{\sum \frac{\# \text{hit} - \text{in} - \text{topK}}{\# \text{citation}}}{N} \quad (3.18)
\]

where k is the total number of candidates patents retrieved, and \#citation is the “ground truth” in the particular query. Then an average of precision and recall of the N queries is computed to represent the performance at the particular k, N is the number of patents in the testing set.
Baseline methods

Since the MeSH ontology similarity is considered to be leverage in the text based information retrieval, a classic method in this field is the cosine similarity [32]. Cosine similarity is the measurement between two word vectors of inner product spaces. The reason it’s called cosine similarity is because the above method is actually measuring the cosine of the angle between the two vectors.

One of the approaches we proposed above is using Indri with MeSH descriptor as queries, the goal is to see whether the MeSH descriptors will serve as query of text representation better than the top-k keywords. As a result, it will be necessary to use the top-k keywords as query for the Indri search engine performance (3.14).

Baseline#1: Cosine similarity, which use the top-30 keywords ranked by tf-idf serving as the vectors.

Baseline#2: Indri search engine with top-30 keywords ranked by tf-idf serving as queries.

Experimental Result

MeSH ontology semantic similarity measure analysis

In Chapter 3, we proposed a three-step calculation for applying MeSH ontology semantic similarity into patent citation recommendation. Specifically, except we only proposed one method for step-2, using AMM, we proposed four different methods of calculation tree node similarity, including two path finding measure Wu&Palmer and Leacoc&Chodorow and two Information content measure Jiang&Conrath and Lin at step-1. We also proposed two ways to measure semantic similarity, the original AMM and modified tf-idf AMM, in step-3.
Here we directly extract all the MeSH descriptors or related entry terms from the claim of each patent document, serve as the assigned MeSH descriptors of each patent. We performed eight similarity measures combining the four different tree node measurements and two document measurements. All the eight tests are performed based on the core patent candidate pool, which contains 28438 patents under Biotechnology category. The 100 patent test set is also generated from the Biotechnology category.

![Figure 4-1 Evaluation of MeSH ontology semantic similarity using different tree node and document similarity measure.](image)

Table 4-1 Precision of eight semantic similarity measure with selected k-value

<table>
<thead>
<tr>
<th>k-value</th>
<th>WP</th>
<th>LC</th>
<th>JC-IC</th>
<th>Lin-IC</th>
<th>WP</th>
<th>LC</th>
<th>JC-IC</th>
<th>Lin-IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0185</td>
<td>0.0195</td>
<td>0.0165</td>
<td>0.014</td>
<td>0.019</td>
<td>0.0145</td>
<td>0.011</td>
<td>0.0075</td>
</tr>
<tr>
<td>40</td>
<td>0.0125</td>
<td>0.0123</td>
<td>0.0095</td>
<td>0.009</td>
<td>0.0143</td>
<td>0.011</td>
<td>0.008</td>
<td>0.0055</td>
</tr>
<tr>
<td>100</td>
<td>0.0069</td>
<td>0.0069</td>
<td>0.0056</td>
<td>0.0051</td>
<td>0.0083</td>
<td>0.006</td>
<td>0.0061</td>
<td>0.0036</td>
</tr>
<tr>
<td>200</td>
<td>0.0046</td>
<td>0.0046</td>
<td>0.0040</td>
<td>0.0035</td>
<td>0.0056</td>
<td>0.0042</td>
<td>0.0045</td>
<td>0.0031</td>
</tr>
<tr>
<td>500</td>
<td>0.0027</td>
<td>0.0024</td>
<td>0.0023</td>
<td>0.0021</td>
<td>0.0031</td>
<td>0.0024</td>
<td>0.0027</td>
<td>0.0019</td>
</tr>
<tr>
<td>1000</td>
<td>0.0016</td>
<td>0.0017</td>
<td>0.0015</td>
<td>0.0014</td>
<td>0.0020</td>
<td>0.0016</td>
<td>0.0018</td>
<td>0.0014</td>
</tr>
<tr>
<td>2000</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0012</td>
<td>0.0010</td>
<td>0.0011</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
Table 4-2 Recall of eight semantic similarity measure with selected k-value

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0759</td>
<td>0.0702</td>
<td>0.0614</td>
<td>0.0502</td>
<td>0.0740</td>
<td>0.0407</td>
<td>0.0272</td>
<td>0.0215</td>
</tr>
<tr>
<td>40</td>
<td>0.2076</td>
<td>0.0908</td>
<td>0.0715</td>
<td>0.06780</td>
<td>0.1204</td>
<td>0.0595</td>
<td>0.0631</td>
<td>0.0270</td>
</tr>
<tr>
<td>100</td>
<td>0.1539</td>
<td>0.1496</td>
<td>0.1089</td>
<td>0.10582</td>
<td>0.1795</td>
<td>0.0808</td>
<td>0.1241</td>
<td>0.0631</td>
</tr>
<tr>
<td>200</td>
<td>0.1908</td>
<td>0.1885</td>
<td>0.1652</td>
<td>0.1430</td>
<td>0.2430</td>
<td>0.1443</td>
<td>0.2100</td>
<td>0.1421</td>
</tr>
<tr>
<td>500</td>
<td>0.2815</td>
<td>0.2448</td>
<td>0.2342</td>
<td>0.2200</td>
<td>0.3236</td>
<td>0.2423</td>
<td>0.2857</td>
<td>0.2046</td>
</tr>
<tr>
<td>1000</td>
<td>0.3271</td>
<td>0.3470</td>
<td>0.3167</td>
<td>0.2814</td>
<td>0.3979</td>
<td>0.3153</td>
<td>0.3714</td>
<td>0.3102</td>
</tr>
<tr>
<td>2000</td>
<td>0.4267</td>
<td>0.4192</td>
<td>0.3978</td>
<td>0.3738</td>
<td>0.5037</td>
<td>0.3941</td>
<td>0.4909</td>
<td>0.4149</td>
</tr>
</tbody>
</table>

Figure 4-1 shows the average recall (Figure 4-1-a) and precision (Figure 4-1-b) of the 100 patents in the test set. The X-axis k-value is the number of top candidates retrieved for recommendation, the k goes up to 2000; while the Y-axis is the average recall/precision of the 100 patents at the particular k. From Figure 4-1-a it is clearly using the path finding measure Wu&Palmer for step-1 with tf-idf AMM in step-3 performs the best at all k. Lin with tf-idf AMM achieves least recall rate at k<600; while with k grows, Lin with tf-idf AMM starts to perform better than Jiang&Conrath with tf-idf AMM and Lin, Jiang&Conrath with original AMM. Table 4-1 and Table 4-2 show recall and precision of selective k-value.

When comparing using the same step-1 method (tree node similarity), except when using Leacock&Chodorow, result of applying tf-idf AMM in step-3 all exceed the performance of applying original AMM (Lin with tf-idf AMM exceed the original AMM only when k>500). While on the other hand, there are no significant patterns for step-1 methods’ performances. As from Figure 4-1, when using original AMM in step-3, Wu& Palmer and Leacock&Chodorow performs very similar and exceed the other two, with Jiang&Conrath to be next and Lin performs the poorest. However, when using tf-idf AMM, Wu&Palmer performs the best with Jiang&Conrath to be the second, Leacock&Chodorow performs slight better than Lin before k reaches 500, and completely falls behind Lin when k exceed 1000.
Particularly in Table 4-1, Leacock&Chodorow with original AMM retrieves highest precision of 0.0195 at k=20, but recall at k=20 still below Wu&Palmer with tf-idf AMM. This is because Leacock&Chodorow with original AMM is performing best at one particular patent with more than 20 references. This result in the fact that this method retrieves a lot absolute number of references so the precision becomes high while recall remains low. Later Wu&Palmer with tf-idf also becomes the best soon after k=40.

In general using modified tf-idf AMM method for step-3 is better than the original AMM, while Wu&Palmer performs best among the four tree node similarity measures (step-1 measure). Even though patent citation recommendation is recall-oriented task, we can see from the average precision, the performance is the same as recall, which also support our result. It is safely to conclude that the combination of Wu&Palmer as step-1 measure with modified tf-idf AMM as step-3 measure would be the best patent document MeSH ontology semantic similarity measure, which is consistent with the actual result. The following measurement set will be applied in all the later analysis: step-1 Wu&Palmer measure; step-2 Average Maximum Measure (AMM, the only proposed step-2 measure); step-3 modified tf-idf Average Maximum Measure.

**Patent candidate selection analysis**

As stated in Chapter 3, we have two potential candidate pool: the core candidate pool of 28438 patents from Biotechnology subcategory under Drug&Medical main category; and the extended candidate pool containing 87193 patents combining the patents from Biotechnology or Drug subcategories under the Drug&Medical main category. Figure 4-2 shows the performance of applying the best method from the last section in two these two pools.
Figure 4-2 Performance of applying Wu&Palmer as step-1 method in patent citation recommendation.

Similar as in last section, the X-axis k-value is the number of top candidates retrieved for evaluation; while Y-axis is the precision/recall rate. For the concern of the potential difference at the scope of the text in Drug and Biotechnology, we evaluate the candidate pool performance with both the original and modified tf-idf AMM for step-3 in the method set.

From Figure 4-2, it is clearly that the performance of using Biotechnology patents exceeds the Biotechnology-Drug combination set with both step-3 measures. It can be inferred that the patents under Drug subcategory may not be adapted to the MeSH ontology well compared with the Biotechnology subcategory. It is understandable as the Drug category may depend more on the terminology related with chemistry and small molecules. That may be the reason UMLS has a separated ontology RxNorm[33], a specific ontology in medicine containing all medication in the US market.

However, one interesting fact is that in combination set, the performance of original AMM in step-3 is actually better than the modified tf-idf AMM. Since the patents under Drug category may not be well adapted to the MeSH ontology, there could be very few MeSH descriptors extracted from the Drug patent, with the frequencies to be low, thus tf-idf may not
well describe the status of MeSH descriptors in the patent, or even the MeSH descriptors can not represent the core concept of the patent, as they may not even be the real important keywords. As a result, the 28438 patents in Biotechnology subcategory will be used in the later analysis.

**MeSH descriptor Assignment Analysis**

One obvious yet very important problem is how to assign proper MeSH descriptor to each patent. Though direct extraction is the most straightforward way to assign the MeSH descriptors to each patent.

![Figure 4-3 Distribution of number of MeSH descriptors assigned to each patent.](image)

Figure 4-3-a shows the distribution of number of MeSH descriptors assigned to each patent document. The MeSH descriptors are directly extracted from the content and are using patents in the Biotechnology pool. In particular, X-axis is the number of MeSH descriptors in one patent document, and Y-axis is the number of The scale of X-axis only gets up to 50, about
500 patents are not included in the figure, which have more than 50 MeSH descriptors assigned to each of those patents. Specifically, the patent having largest MeSH descriptor set possesses 211 descriptors. On the other hand, as shown in the figure, 231 patents do not have any MeSH descriptors assigned. This extreme unbalanced MeSH descriptor assignment may lead to the bias in the semantic similarity measure, as the more MeSH descriptors one patent possessed, the higher semantic similarity score it potentially can achieve.

The MeSH ontology is originally developed to index the journal articles in the PubMed database under the National Library of Medicine, which most time by now is still manually assigned, but there are a lot of programs developed to assist the Indexers to select the proper MeSH descriptors. Considering the complexity and difficulty, it is impossible to have expert manually assign the MeSH descriptors to the patent documents, Medical Text Indexer (MTI), which is the main program employed by NLM for MeSH indexing, is applied here for MeSH descriptor auto assignment.

Figure 4-3-b shows the distribution of MeSH descriptors in each patent using MTI auto assignment. This assignment is applied to the same Biotechnology patent pool. Comparing with Figure 4-3-a, the distribution is more concentrated, all the patents have at least one MeSH term assigned, and only 178 patents got one MeSH descriptors assigned. Moreover, we set the maximum number of MeSH descriptors assigned to a patent to be 50, so all the 28438 patents are in the plot. Comparing the Figure 4-3-a which has a mean of 15 and standard deviation of 11.5, Figure 4-3-b has mean of 10 with standard deviation of 4.3. Even though the number of MeSH descriptors in one patent is still not consistent across the whole patent pool, the variation is for less than in Figure 4-3-a.
Figure 4-4 Performance of using MeSH descriptors from Direct extraction and MTI auto assignment.

We further applied the two MeSH descriptor assignments in the Biotechnology patent set, and use the “best method set” to evaluate the semantic similarity performance. Similar as previous sections, average recall and precision for the top-k retrieved candidates are computed as the evaluation metrics, as shown in Figure 4-4. The recall rate is very similar at k<180, then The MTI assignment retrieved a much higher recall rate afterwards. Also when using the direct extraction, the top candidates did not hit any “true references” until k reaches 40. The precision rate also supports this conclusion.

MeSH descriptor in content-based similarity measure analysis

In the last section of Chapter 3, we proposed a method by using both MeSH descriptors and top-k keywords by tf-idf as query terms for a content-based search engine Indri[30]. We proposed the combination query $Q_c = (1-\alpha) \cdot Q_{cnt} + \alpha \cdot Q_{MeSH}$, which $Q_{cnt}$ is using top-30 keywords and $Q_{MeSH}$ is using MeSH as query; $\alpha$ is the weight for the content query. When $\alpha=0$, the query
becomes pure top-30 keywords; and $\alpha=1$ turns into pure MeSH descriptor based query. The MeSH descriptor is assigned using MTI program, and using the Biotechnology patent pool as the candidates.

Table 4-3 The average recall rate of top-k candidates for the combination queries.

<table>
<thead>
<tr>
<th>k-value</th>
<th>Top-30</th>
<th>$\alpha=0.2$</th>
<th>$\alpha=0.3$</th>
<th>$\alpha=0.4$</th>
<th>MeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.2513</td>
<td>0.2547</td>
<td>0.2607</td>
<td>0.2672</td>
<td>0.1654</td>
</tr>
<tr>
<td>100</td>
<td>0.4329</td>
<td>0.4181</td>
<td>0.4160</td>
<td>0.4133</td>
<td>0.2992</td>
</tr>
<tr>
<td>500</td>
<td>0.6217</td>
<td>0.6346</td>
<td>0.6294</td>
<td>0.6279</td>
<td>0.4563</td>
</tr>
<tr>
<td>1000</td>
<td>0.6792</td>
<td>0.6995</td>
<td>0.6967</td>
<td>0.6938</td>
<td>0.6047</td>
</tr>
<tr>
<td>5000</td>
<td>0.7906</td>
<td>0.8198</td>
<td>0.8287</td>
<td>0.8344</td>
<td>0.7343</td>
</tr>
<tr>
<td>10000</td>
<td>0.8395</td>
<td>0.8636</td>
<td>0.8581</td>
<td>0.8537</td>
<td>0.7827</td>
</tr>
</tbody>
</table>

Figure 4-5 Evaluation of applying MeSH descriptors and top-k keywords in content-based similarity measure.

Figure 4-5 shows the average recall/precision rate of this combination queries. It is interesting to see that the pure MeSH descriptor query actually retrieves lowest recall rate at all the k-values. It is understandable as the mean of the MeSH descriptor numbers in a patent is only 10, comparing to 30 tf-idf keywords are applied in the query. And for content-based similarity measure, the number of the elements in one query can affect the similarity computation.
A few of the combination queries achieve highest at different k-value. Table 4-3 selects some of the recall rate at different k, including the pure MeSH, pure top-30 keywords queries, and some of the combination queries in different α. The queries retrieve highest when α=0.2, 0.3, 0.4, each retrieves highest at some level of k. Since for the search engine Indri, when using a combination query, the similarity score for the Q_MeSH and Q_cnt is computed individually and only combined at the last step, so the number of elements in the query should not bias the similarity measure. This suggests the MeSH descriptors can help increase the similarity measure at the content-based level, but not too much, this could be the reason that the frequency of technical terms in the document is usually low.

Baseline comparison

From the previous sections, we have generated a “best semantic similarity measure set” by using MTI MeSH assignment, Wu&Palmer tree node similarity and modified tf-idf AMM document similarity, and applied into the Biotechnology patent pool. We evaluated this semantic similarity measure with the two baseline methods.

Figure 4-6 Performance evaluation: MeSH semantics and baseline comparison
We compare the pure MeSH ontology semantic similarity with the two baselines: 1> Indri search engine using top-30 tf-idf keywords as query and 2> cosine similarity measure. We put a third method generated above by combing the top-30 keywords and assigned MeSH descriptors as query, using $\alpha=0.4$ ($\alpha$ is the weight of MeSH descriptor query), which is one of the best keywords-MeSH combination Indri query performance. Figure 4-6 shows the recall and precision as the evaluation of the above four methods. As shown in the figure, baseline#2, the cosine similarity achieves lowest recall rate with 0.4 at $k=1000$, MeSH semantic similarity is about consistently 20% higher with recall rate of 0.54 at $k=1000$. The two Indri performances achieve very similar result, and yet the best among those four methods. Before $k=240$ they behave very similar, while $k@240$~480 top-30 keywords query achieve higher recall rate; while after that the combination query keeps achieving the highest recall rate. Although Indri performing better MeSH semantic similarity suggests when purely depends on the semantics, MeSH ontology may not be as good as a lot other sophisticated information retrieval engine, it’s better performance than cosine similarity still suggests there’s potential to integrate MeSH semantics into other recommendation system may further improve the performance. Moreover, the combination query’s performance suggests MeSH vocabulary as technical terminology, may also play a special role more important than the usual tf-idf based keywords.
Chapter 5
Discussion

Baseline comparison

From the previous section, we can see that pure MeSH ontology semantic similarity actually is not performing better than some sophisticated content-based information retrieval engine, such as Indri. Part of the reason could be because the MeSH indexing itself may not be a strong support for information retrieval. MeSH ontology is primarily developed for the indexing of PubMed articles. In the search and recommendation algorithm of PubMed database

Another interesting fact in the experiment when comparing the MeSH semantic similarity with the Indri top-30 keywords recommendation result is that, although the recall rate for MeSH similarity is lower than Indri, for a lot of the items in the testing set, MeSH similarity actually retrieves “true reference” ranking very low at Indri at a much higher ranking.

Table 5-1 shows one particular example of rankings at MeSH semantic similarity and Indri top-30 keywords query. Biotechnology patent “7217537” is entitled “Method to increase carotenoid production in a microbial host cell by down-regulating glycogen synthase”, it has 48 MeSH descriptors in total, which means at least in this particular case, the number of terms used is not the main effect leading to a higher recall rate at k=5000. However, this patent is using

<table>
<thead>
<tr>
<th>True references</th>
<th>5182208</th>
<th>5466599</th>
<th>5935808</th>
<th>5972642</th>
<th>6015684</th>
<th>5691190</th>
<th>6825002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indri: top-30 keywords</td>
<td>40</td>
<td>180</td>
<td>240</td>
<td>480</td>
<td>2400</td>
<td>4820</td>
<td>N/A</td>
</tr>
<tr>
<td>MeSH semantics</td>
<td>320</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>840</td>
<td>1640</td>
</tr>
</tbody>
</table>
more terms in MeSH semantic similarity than in the Indri query, the MeSH semantic similarity is
retrieving fewer “true references”. On one hand, it seems to indicate the power of MeSH
semantic similarity in citation recommendation is weaker than Indri; on the other hand, reference
“5691190” only ranked 4820 when using Indri for recommendation while ranked much higher at
840 when using MeSH semantic similarity; reference “6825002” is not appearing in top 5000
when using Indri but ranked 1640 when using MeSH semantic similarity. This suggests the
MeSH semantics may captured some technical features that can not represented by keywords but
rather represented by some low frequency technical terms, the rankings of the candidate pool
using two recommendation methods result in different structure, some of the references that may
not be able to captured in Indri (or some other content-based methods) can be captured by the
MeSH semantic similarity.

As a result, though the MeSH semantic similarity may not be a good recommendation
engine by itself, it could play the role as an extra feature of a lot other popular content-based
citation recommendation method.

**MeSH descriptor assignment**

As stated, the MeSH ontology is primarily designed for the PubMed article indexing, and
seldom used outside the MEDLINE database. For most articles, the MeSH descriptors are
manually or semi-manually assigned. MeSH Terms can usually be found below the citation when
at abstract format. Figure 5-1 shows an example of a PubMed records with assigned MeSH
records. Typically one article will have ten to twelve MeSH terms. Most MeSH terms are Main
Headings, which are selected from the MeSH descriptors; those listed in the same line as Main
Headings but after the forward slash is called subheadings, which are selected from the MeSH
qualifier set. MeSH descriptors indicate the subjects of the article, making up most of the MeSH
vocabulary and usually updated annually. MeSH qualifier is a much smaller yet stable set, which usually describe the genre rather than the content and seldom updated. For example, in PMID: 20373513 for the MeSH term “Gene Expression Profiling/method”, “Gene Expression Profiling” is the Main Heading from the MeSH descriptor set; which indicates the article is using the gene expression profiling technic for analysis, while the subheading “method” is just a further specification indicates the Main Heading is a kind of method. Judging from the descriptor/qualifier set definitions, we only apply the descriptors in our experiment.

![Figure 5-1 Example of MeSH terms in a PubMed record with PMID: 20373513](image)

As indicated, MeSH terms in PubMed are assigned manually by the Indexing experts with the assistance of various program like MTI. While in our experiment, all the MeSH descriptors are assigned by programs without human intervention, resulting in a very uneven distribution of the MeSH descriptor numbers. Comparing with a very restricted scale of 5~15 MeSH terms in the PubMed articles, a very significant number of patent documents have more than 30 MeSH descriptors even with MTI assignment, this could result in redundant or unrelated information, resulting in noises in similarity measure.

Further, the MeSH ontology is mainly applied for the indexing and cataloging in the MEDLINE database. As a result, the major role for MeSH in PubMed search is direct mapping
with queries or related variants. The MeSH semantics is mainly applied in the MeSH database search, which is used for helping users to choose the most appropriate MeSH terms candidates from the MeSH vocabulary, which will later used in the main PubMed database search. In other words, the MeSH semantics is not directly applied in the PubMed article search, and the indexing mapping is often combined with other metadata, like title or keywords search.

In MEDLINE database, MeSH terms are manually assigned or at least evaluated by the indexing experts and the number of terms are restricted to 5~15. This reduces the bias of semantic similarity measure of MeSH ontology in most of the researches focused on the MEDLINE database [13][14][22].

Some other studies tried to assign the MeSH ontology to short text like image annotation either by using MTI or direct extraction [15]. Our study is also one of the pioneer researches on assigning MeSH descriptors to full text. Similar as MEDLINE database, which uses abstract as the background text for MeSH term assignment, we use claims instead, as claims are consider to be the most representative element in a patent document, and has been widely accepted to be used for query generation. However, even we only use the claims, the distribution of MeSH descriptor numbers still vary a lot according to Figure 4-3.

The above fact indicates the MeSH vocabulary is well developed for describing the content of biomedical documents, but the MeSH semantics may not be strong enough to act as an individual engine or a major feature in search engine. This is further approved by our result that the MeSH semantic similarity measure is retrieving lower recall rate than the comprehensive search engine Indri.

However, similar as PubMed articles, with proper assignment, MeSH ontology should still able to be applied as one feature to help improve the search performance, when combining with other metadata, such as title or keywords in the PubMed search.
Information Loss

As described in Chapter 2, only 8 out of 16 trees are included in the study. The other 8 trees are considered to be not strictly related to the biomedical domain. However, whether to use subtrees like Information Sciences [L] and Health Care [N] should be further discussed. Currently these two subtrees are not included in the study, since they mostly contain information of other domains, even the subtree N is called Health Care, it is mostly provide information about policy, Economics and Administration. These were considered to provide more noises than useful information in the study. However, it should be realized that part of the subtrees should be eligible to be included in the study, since for information like Health Care Assessment and Facility may provide extra insight if two patent share the same these kind of MeSH descriptors.

Another concern about information loss is the simplification of the three-layer structure of MeSH descriptor. As stated in Chapter 2, one descriptor usually contains more than one concepts, one of the concepts is selected as the “preferred concept”, other concepts are more or less describing similar topic but slightly different scope, and annotated as “wider” or “narrow” than the preferred concepts. Similar is concept to term, which means one concept usually has a set of terms. In the study we have ignored the concept phase and treat all the terms in one descriptor under all concepts the same, as the synonyms of this descriptor. However, this concept level, particularly the evaluation of scope difference between each other should be further defined and integrated into the similarity measure.
Chapter 6
Conclusion

In this study, we examined the MeSH descriptor semantic similarity measure for patent citation recommendation. The experimental results generate the “best measure set” for the three-step measurement: 1> Wu&Palmer path finding measure performs best among all the eight measurements for the tree-node similarity; 2> The Average Maximum Measure (AMM) is applied for the MeSH descriptor similarity measure and 3> A modified AMM which utilizes the tf-idf information for each of the MeSH descriptors in the patent document is selected for evaluating the MeSH semantic similarity between two documents.

Particularly, information of tf-idf is having a major impact during the MeSH semantic similarity, as for different measures of MeSH semantics, integrating tf-idf information of MeSH descriptor in document similarity measure will improve the result.

We also evaluated our “best measure set” with two baseline methods: cosine similarity and Indri content-based search engine. The two baseline methods use the top-30 tf-idf keywords as the query. Results indicates that the average recall rate of our method outperform the cosine similarity by 20% when retrieving the top 1000 candidates, but the Indri search engine is performing better than ours, which indicating MeSH semantics, though may not be a strong relation for the patent document retrieval, may still act as one helpful feature to be incorporated into more complex recommendation system.

In the end, we tested on the performance of using MeSH descriptors assigned in different ways. Again by using our “best measure set”, results shows using MeSH descriptors assigned by the Medical Text Indexer (MTI), the average recall rate outperformed the directed extract MeSH descriptors by 34%.
Further work could include expanding this MeSH assignment and similarity measure by incorporating other major medical ontologies such as SNOMED-CT [34]. Additionally, MeSH descriptors can be incorporated into other major recommendation system as an extra feature. Moreover, the concept level and definition of the scope in the descriptor structure should be further be integrated into similarity measure.
Bibliography

16. Taschwer, M. Text-Based Medical Case Retrieval Using MeSH Ontology.


34. SNODMED Clinical Terms http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
Appendix

MeSH ontology

The Medical Subject Headings (MeSH) thesaurus is updated annually, the dataset we used here is the 2013 version. The XML format MeSH records are downloaded from the NLM website. A complete MeSH structure contains 16 subtrees: Anatomy [A], Organisms [B], Diseases [C], Chemicals and Drugs [D], Analytical, Diagnostic and Therapeutic Techniques and Equipment [E], Psychiatry and Psychology [F], Phenomena and Processes [G], Disciplines and Occupations [H], Anthropology, Education, Sociology and Social Phenomena [I], Technology, Industry, Agriculture [J], Humanities [K], Information Science [L], Named Groups [M], Health Care [N], Publication Characteristics [V] and Geographics [Z]. To restrict the terminology into strictly biomedical related, we only select the vocabulary with related structure of nine subtrees A-H and J. More detailed details of each subtree is described at the MeSH website https://www.nlm.nih.gov/mesh/trees.html.

Each of the MeSH record in the XML belongs to one of the following type: Descriptor, Qualifier and Supplementary Concept Record (SCR). Descriptor is the dominant type in the whole MeSH ontology, which is used to describe the content of the MEDLINE articles; 183 Qualifiers are contained in the MeSH ontology 2013 version, qualifiers are general terms to define the type and other metadata of the article; SCR are created daily without further rigorous curation, most of the SCRs are alternative names of chemicals or drugs or related instance of reaction or literature that describing the main record. For this reason, only descriptors are included in the research.
Implementing the semantic similarity measure

The three-step semantic similarity measure is implemented in Python 2.7.3. Specifically, for the MeSH direct extraction, the tokenization and other preprocess of the patent documents are implemented using the python NLTK (Natural Language Toolkit) 3.0 package. Since the entry terms of the MeSH descriptor contain the morphological forms of the descriptor, no stemming process is applied. The cosine similarity is implemented by the python package gensim 0.8.6.

The MTI MeSH descriptor auto assignment is using the Batch Medical Text Indexer, which requires a UMLS account. The online portal is at http://skr.nlm.nih.gov/batch-mode/mti.shtml. Claim of each patents are uploaded to the batch, all the settings are set to be default except the following: Basic Filtering Options are set to be “Medium”; “Use MTI as First Line Indexer (MTIFL)” is checked at Filtering Modifiers; Individual Item Timeout is set to be “30min”; “show tree code” is checked besides all the default settings at Post Processing. It could take days to process all the 28438 items depending on the workload of the batch. In the output, only descriptors are kept; if one item is assigned with more than 50 descriptors, only the first 50 are accepted based on the final score.

The Indri search engine which performs as one of the baselines, is indexed and run with default option. The content-based Indri search uses the following query:

\[ \text{#weight(tf-idf}_1 \text{ wd}_1 \text{ tf-idf}_2 \text{ wd}_2 \ldots \text{ tf-idf}_n \text{ wd}_n) \], where \text{wd}_i is the ith word in the top-30 keywords and \text{tf-idf}_i is its tf-idf in the corpus.

Further when applying the Indri in the the combination of top-30 keywords and MeSH descriptors for content-based retrieval usage, the query is as following:

\[ \text{#weight( #weight(\alpha \text{ #weight(tf-idf}_1 \text{ wd}_1 \text{ tf-idf}_2 \text{ wd}_2 \ldots \text{ tf-idf}_n \text{ wd}_n)) \text{ #weight((1- \alpha) \text{ #weight(max(tf-idf) syn(MH}_{11}, \text{ MH}_{12}, \ldots, \text{ MH}_{1j}) \ldots max(tf-idf) syn(MH}_{n1}, \text{ MH}_{n2}, \ldots, \text{ MH}_{nj}))))),} \] where the \( \alpha \) is the weight for the whole top-30 keywords query, and max(tf-idf) is the highest tf-idf in the document, \( \text{MH}_0 \) is the jth entry terms in the MeSH descriptors.