

EXPLOITING SEMANTICS FOR IMPROVING CLINICAL INFORMATION RETRIEVAL

Atanaz Babashzadeh

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO
2013

© Atanaz Babashzadeh, 2013

Abstract

Clinical information retrieval (IR) presents several challenges including terminology mismatch and granularity mismatch. One of the main objectives in clinical IR is to fill the semantic gap among the queries and documents and going beyond keywords matching. To address these issues, in this study we attempt to use semantic information to improve the performance of clinical IR systems by representing queries in an expressive and meaningful context. In this study we propose query context modeling to improve the effectiveness of clinical IR systems. To model query contexts we propose two novel approaches to modeling medical query contexts. The first approach concerns modeling medical query contexts based on mining semantic-based AR for improving clinical text retrieval. The query context is derived from the rules that cover the query and then weighted according to their semantic relatedness to the query concepts. In our second approach we model a representative query context by developing query domain ontology. To develop query domain ontology we extract all the concepts that have semantic relationship with the query concept(s) in UMLS ontologies. Query context represents concepts extracted from query domain ontology and weighted according to their semantic relatedness to the query concept(s). The query context is then exploited in the patient

records query expansion and re-ranking for improving clinical retrieval performance. We evaluate this approach on the TREC Medical Records dataset. Results show that our proposed approach significantly improves the retrieval performance compare to classic keyword-based IR model.

Acknowledgements

I would like to express my deepest appreciation to my supervisor Professor Xiangji (Jimmy) Huang for his support, motivation, guidance, enthusiasm, leadership and enormous knowledge. I would like to thank him for his continuous support, encouragement and engagement through the learning process of this master thesis. Without his guidance and persistent help this dissertation would not have been possible.

I would also like to thank my thesis committee: Prof. Ali Asgary and Prof. Nick Cercone for their insightful comments and suggestions.

In addition I would like to show my greatest appreciation to Dr. Mariam Daoud for her tremendous support and help and who has been instrumental in the successful completion of this thesis. Sincere thanks to my lab mates in the Information Retrieval and Knowledge Management lab whose help, time and support has helped me through the challenges.

Finally, I am deeply grateful to my family for their enormous and continuous support, this work would not be able to come together without them.

Table of Contents

1. Introduction	1
1.1. Background	1
1.2. Motivation.....	2
1.3. Contributions	3
1.4. Thesis Structure	5
2. Literature Review	7
2.1. Semantic Information Retrieval.....	7
2.2. Association Rule Mining in Clinical Retrieval.....	9
2.3. Information Retrieval by Semantic Similarity.....	12
2.4. Medical TREC Records Track Participations	13
3. Semantic Information Retrieval.....	16
3.1. Unified Medical Language System (UMLS).....	17
3.1.1. Metathesaurus.....	18
3.1.2. Semantic Network	18
3.1.3. SPECIALIST Lexicon.....	19
3.2. Conceptual Representation	19
3.2.1. MetaMap.....	20
3.2.2. Conceptual Indexing.....	20
3.3. Query Domain Ontology	23
3.4. Query Concepts Extraction using Query Domain Ontology	25
4. Association Rule Mining in Clinical Retrieval.....	27
4.1. Association Rule Mining	28
4.2. Association Rule Mining in Information Retrieval	30
4.3. TopKRules Algorithm	32
4.4. Query Concepts Extraction using Association Rule Mining	34
5. Information Retrieval by Semantic Similarity and Semantic Relatedness	37
5.1. Semantic Similarity and Semantic Relatedness Measures	38
5.1.1. Semantic Similarity	38
5.1.1.1. Path Lengths.....	39
5.1.1.2. Wu and Palmer.....	39
5.1.1.3. Leacock and Chodorow.....	40
5.1.1.4. Hirst and St-Onge.....	40
5.1.2. Semantic relatedness	40
5.1.2.1. Resnik 41	
5.1.2.2. Jiang and Conrath.....	42
5.1.2.3. Gloss Vector.....	42

5.2. Semantic Query Context Modeling	44
5.2.1. Query Concepts Weighting	45
5.2.2. Semantic Query Context.....	46
5.2.2.1. Modeling Query Context Using Query Domain Ontology	46
5.2.2.2. Modeling Query Context Using AR Mining.....	47
5.3. Semantic Query Expansion.....	47
5.4. Semantic Re-ranking	48
6. Experimental Settings	50
6.1. Information Retrieval System Settings	50
6.1.1. Pre-processing Settings	51
6.1.2. Indexing Settings	51
6.1.3. Retrieval Settings.....	52
6.1.4. Query Expansion Settings	54
6.2. UMLS Settings	55
6.2.1. UMLS::Interface.....	55
6.2.2. UMLS::Similarity.....	56
6.3. Dataset.....	56
6.3.1. Patients Records	58
6.3.2. Query Set	60
6.3.3. Retrieval task.....	61
6.3.4. Relevance Judgments	61
6.3.5. Evaluation Measures.....	62
6.4. Parameter Tuning.....	64
6.5. Experimental Runs	67
6.5.1. Baseline Runs	68
6.5.2. Query Expansion Runs	69
6.5.3. Semantic Query Context Runs	70
6.5.3.1. Association Rule Mining Runs	70
6.5.3.2. Query Domain Ontology Runs.....	72
7. Experimental Results	74
7.1. Performance of Semantic Query Context Based on AR Mining.....	74
7.2. Performance of Semantic Query Context Based on Query Domain Ontology	75
8. Analysis and Discussion	78
8.1. Impact of Conceptual Representation on Retrieval Performance	78
8.2. Impact of Incorporating Keywords Dependency in Retrieval Framework	82
8.3. Impact of Semantic Query Context on Retrieval Performance.....	84
8.3.1. Impact of Semantic Association Rule Mining on Retrieval Performance.....	84
8.3.1.1. Impact of AR Mining on Query Context Modeling	84
8.3.1.2. Impact of Semantic Relatedness on Query Context Modeling.....	85
8.3.1.3. Impact of Using AR Mining and Semantic Relatedness for Query Context Modeling.....	86

8.3.2.	Impact of Query Domain Ontology on Retrieval Performance.....	88
8.3.2.1.	Impact of Query Expansion Using Semantic Query Context.....	88
8.3.2.2.	Impact of Re-ranking Using Semantic Query Context.....	89
8.3.3.	Impact of Combining Query Expansion and Re-ranking Using Semantic Query Context 90	
8.3.4.	Impact of Rocchio’s Query Expansion Mechanism.....	91
8.3.5.	Impact of integrating Rocchio into Semantic Query Context	92
8.3.6.	Significance test.....	93
8.4.	Comparison of Approaches.....	99
9.	Conclusions and Future Work	102
9.1.	Semantic Association Rule Mining for Improving Clinical Text Retrieval	102
9.2.	Semantic Query Domain Ontology for Improving Clinical Text Retrieval	103
9.3.	Conclusions.....	104
9.4.	Future Work.....	106
	Works Cited.....	108
A.	IR System’s Settings.....	115
A.1	Data Pre-processing	115
A.2	Building Index	116
A.3	Documents Retrieval.....	116
A.4	Terrier’s Properties	117
B.	Conceptual Mapping Settings	119
C.	Mapping Reports to Visits	120
D.	Re-ranking Script.....	124
E.	TOPKRULES Algorithm.....	127
F.	Perl Program for Modeling Query Domain Ontology	129
G.	Perl Program for Calculating Semantic Relatedness.....	131
H.	Evaluation Scripts.....	139
I.	Topics of the TREC Medical Records Track.....	140
J.	Sample Raw data	142

1. Introduction

1.1. Background

Due to the escalating quantity of digitalized medical patient records, the need for advanced information retrieval (IR) and knowledge discovery systems increases. Digitalized medical patient records contain valuable information such as the current clinical diagnosis, medical history, medications, allergies, examination findings, results of investigations, nursing observations, treatment plans and notes from other allied health professionals. This information is usually embedded in medical reports in the form of text, discharge summaries and progress notes. Therefore it is difficult for clinicians to rapidly access the desired content of a record and the need for effective clinical IR systems arises. Clinical IR systems improve access to updated clinical information in compare with traditional sources of information i.e. colleagues, notes, printed textbooks or journals [45], they also offer advantages for resolving clinical crisis, questions and problems, accessing information needs in timely manner, assisting in decision-making

process, dealing with memory issues, and fulfilling educational objectives [45].

In medical and clinical IR systems, the desired information of a query or topic requested by clinicians is usually a list of entities or records covering different aspects that are related to the query. In a general context, the various aspects could cover diseases, symptoms, treatments, etc. Thus it is important for a medical IR system to be able to provide comprehensive and diverse answers to fulfill clinicians' requests. As result a patient may derive a right treatment after quick and meaningful diagnosis of disease and treatment suggestion. This helps not only the patients in receiving the good health but also the doctors in keeping their good will [45].

1.2. Motivation

IR in medical and clinical context presents several problems including terminology mismatch due to the frequent use of acronyms, homonyms, synonyms and granularity mismatch due to the use of subsumed concepts and semantic gap among queries and documents. Semantic gap illustrates the difference among different descriptions of an object in different representations, which may result in uncommon understanding of the requester and provider [17]. One of the main objectives of clinical IR is to fill the semantic gap among the queries and documents and going beyond keywords matching. Moreover, current clinical IR approaches concentrate on lexical matching and disregard semantically related and similar concepts. The absence of mutual terms in a query and set of documents doesn't essentially imply that those documents are not related to the query.

Query terms and documents terms can be semantically related while lexically different; thus, classical IR methods fail to retrieve semantically similar documents [21]. It is fairly easy for humans to understand exactly semantically similar terms or semantically related terms, but difficulties arise when we need IR systems to determine semantically similar and closely related terms. Determining semantically similar and semantically related terms can improve clinical IR systems by determining closely related diseases, treatments and symptoms. These problems have raised new challenges in the field of clinical IR, especially given that medical documents and electronic medical records (EHR) are expanding and there is a growing demand for text-based clinical IR systems.

This study lies on the framework of the TREC medical record challenge. The goal of this track is to develop a clinical IR system that given a specific query it can search through EHRs and retrieve a list of relevant documents. Queries specify a particular disease/condition set and a particular treatment/intervention set. The collection consists of a set of de-identified free-text EHRs. The main objective of this study is to improve the retrieval performance of content-based clinical IR systems.

1.3. Contributions

In this study, we propose a novel IR approach that is able to tackle the presented challenges of clinical IR such as terminology mismatch and granularity mismatch and to capture those relevant documents not containing queries' keywords explicitly. The proposed model is capable of combining search by meaning and search by keyword; it

also attempts to fulfill the objective of filling semantic gap between queries and documents. This research presents a novel medical query context modeling to improve clinical text retrieval performance.

The key unique contributions in this research concern (1) modeling query domain ontologies to extract representative concepts to information need, (2) semantic-based AR mining for modeling a medical query context and (3) a semantic-based concept weighting schema to weight concepts semantically in relation to the query.

One of the characteristics of semantic search is conceptual representation rather than simple keywords representation to enhance the representation of queries and documents. Modeling and developing domain ontologies is a fundamental framework for representing knowledge using a set of concepts and the relationships among the concepts. Medical domain knowledge is developed by several different ontologies including Unified Medical Language System (UMLS). In this approach we develop query domain ontology for each query using UMLS ontologies. Query domain ontology consists of all the concepts that have semantic relationship with the query concept/s in UMLS ontologies. In order to identify the extent of a concept relevancy in relation with query concept(s) we weight each concept in query domain ontology according to its semantic relatedness/similarity to the query concept. Query context represents all the concepts in the query domain ontology that are weighted according to their relatedness/similarity to the original query concepts. The query context is used to enhance classical IR model and to calculate and assign a new score to documents by considering semantics.

In this study, we also focus on the use of Association Rule (AR) mining for improving clinical text retrieval. In the biomedical field, association rule mining presents a promising technique for finding hidden patterns in a medical dataset and for improving the information retrieval performance via query expansion. Its objective is to discover all co-occurrence relationships, called associations, among data items. Association rule mining has been widely applied on medical data for diagnosis or symptom prediction, gene expression and cell type prediction, classification purposes and information retrieval. The usefulness of AR mining technique is strongly limited by the huge amount and the low quality of delivered rules. Current data mining techniques can efficiently generate association rules that are statistically significant to the source dataset but these techniques have limited semantic capabilities. The use of AR in IR is challenging since most of the generated rules are trivial, redundant, semantically wrong and conflict with common sense or basic domain knowledge, or already known by end-users. In this research, we propose the use of semantic relatedness/similarity measures using the UMLS semantic network for selecting concepts from rules that semantically cover the query. The selected rules are used to form a medical query context where concepts are weighted according to their relatedness/similarity to the original query concepts. The query context is then exploited in the patient records query expansion and re-ranking for improving clinical retrieval performance.

1.4. Thesis Structure

This thesis is organized as follows: Chapter 2 describes background information and

related work found upon semantic information retrieval, association rule mining in clinical retrieval and information retrieval by semantic similarity and semantic relatedness, this chapter also presents previous medical TREC records track participations. Chapter 3 describes our proposed method for semantic information retrieval; Chapter 4 describes our proposed AR mining algorithm for clinical text retrieval and chapter 5 presents our proposed technique for information retrieval using semantic similarity and semantic relatedness. Chapter 6 describes our experimental settings including system's settings, dataset, parameter tuning and experimental runs. Chapter 7 presents our experimental results and chapter 8 provides a discussion and analysis on our proposed methods and summarizes the contributions of this research. Finally, chapter 9 concludes this study and presents future work of this thesis.

2. Literature Review

Several approaches have been proposed and discussed on the following topics: semantic information retrieval, association rule mining in clinical retrieval and information retrieval by semantic similarity. In this chapter we focus on introducing and summarizing the previous work on these topics and we present our own unique contributions on these topics. This chapter also introduces some of the previous participations in TREC Medical Record challenge.

2.1. Semantic Information Retrieval

One of the main objectives in IR is filling the semantic gap between queries and documents and going beyond keywords matching. One of the characteristics of semantic search is conceptual representation rather than simple keywords representation to enhance the representation of queries and documents. Keyword-based information retrieval models could return imprecise and erroneous results due to their dependency on specific keywords. Concept-based representation approaches are intended to overcome this limitation by representing queries and documents with semantic concepts using

biomedical resources such as UMLS. Although in [19], authors reported that using concept-based representation is not effective and degrades the retrieval performance. However authors in [53] demonstrated that these approaches should be combined with term-based approaches for better performance.

In [30] authors tackle the issue of terminology mismatch by proposing a technique to represent queries with only medical concepts crucial to information need. In their method, they represent queries and documents using medical concepts that are directly related to symptoms, diagnosis, diagnosis test and treatment. Their approach outperforms traditional term-based approach and concept-based baseline where all concepts are present.

Terms in queries are sometimes too general however those terms in relevant documents are too specific or vice versa, this leads to granularity mismatch. In [62] authors tackle the issue of granularity mismatch by exploiting subsumption relationships extracted from medical domain knowledge resources, where subsumption relationships are defined as parent-child relationships when one concept is a subset of another concept. Their results demonstrate potential increase in retrieval performance when considering their proposed approach.

In this study we tackle the problem of terminology mismatch by presenting documents and queries using UMLS medical concepts. We map all the queries keywords and documents keywords to UMLS concepts using MetaMap [5]. Therefore keywords with the same meaning are mapped to the same concept. We also tackle the problem of granularity mismatch by developing query domain ontology for each query using UMLS

Metathesaurus. Query domain ontology consists of those concepts that have semantic relationship with the original query concepts such as those that share some common property or are related by definition.

2.2. Association Rule Mining in Clinical Retrieval

Mining association rules between sets of items in large databases was first introduced in 1993 by Agrawal et al. [1]. Given a large database of customer transactions they presented an efficient algorithm that discovers all significant association rules denoted by $x \rightarrow y$ between items in transactions, where each transaction consists of items purchased by a customer in a visit. x and y are called the antecedent and consequent of the rule respectively.

One of the classic association rule mining algorithms is the Apriori algorithm. The Apriori algorithm is an algorithm for mining frequent itemsets; it attempts to find frequent itemsets that have minimum support where any subset of a frequent itemset must also be a frequent itemset [2]. Given a transaction database and support threshold, standard association rule algorithm has two phases; first phase finds all itemsets satisfying minimum support, it starts with generating frequent one-itemsets and proceeds to two-itemsets and so on, until there are no more frequent itemsets. Second phase generates all rules with support and confidence above specified threshold. Based on the choice of the minimum support and minimum confidence, Apriori algorithm can have the disadvantage of being very slow and producing huge or insignificant amount of information therefore

most of the discovered rules are not useful since they may contain redundant information, irrelevant information or they may illustrate insignificant knowledge.

To address these issues TopKRules algorithm has been introduced. TopKRules algorithm [14] mines the top-k association rules; k is specified by users and represents number of association rules to be found. Experimental results show that TopKRules algorithm has excellent performance and scalability where the number of generated rules can be controlled. Another advantage of this algorithm is that it mines the top-k rules that meet specified confidence, since minsup is more difficult to assign due to its dependency on database characteristics and nature.

One of the main issues in association rule mining is the irrelevancy of most of discovered rules. In [39], authors addressed this limitation by applying search constraints to find only medically significant association rules. The rules were then used to predict heart disease related attributes.

In [54], authors adopted concept lattice and Hasse diagram to mine interesting association rules in checkup medical data. Concept lattices have application in knowledge representation, data analysis, information retrieval and rule extraction. Hasse diagrams of concept lattices generate a hierarchical structure with mapping sequence from more specific lattices to more general lattices, from which interesting association rules can be extracted. In this paper, authors succeeded to discover some interesting association rules between hypertension and some other diseases using Hasse diagrams. Their results show that the use of hierarchical structure of concept lattices is effective in medical data

mining.

In [27], authors proposed a flexible association rule mining technique based on term co-occurrence. This method adds more control parameters to extract interesting association rules by taking into account both the co-occurrence frequency and the confidence and direction of the association rules. The association rules were then used for automatic query expansion and results confirm the effectiveness of their proposed approach. Weighted association rule mining [55] is an extension of traditional association rule mining by assigning a weight to each item to represent their interest/intensity. One of the challenges in the weighted association rule mining is the weight assignment. In [57], authors proposed a self-assigned weighting technique as opposed to user-specified weighting to extract positive and negative association rules. Typical or positive association rules consider only items that appear in the dataset frequently but negative association rules consider negated or absent items, negative association rules are valuable since they can identify items that conflict [4]. To derive the weights for each item, an extension of Kleinerg's HITS model [36] is exploited; weights are then used to extract positive and negative association rules. Their results illustrate that this framework is useful in eliminating many irrelevant, redundant and trivial rules.

One of the limitations of association rule mining in medical domain is that association rules are extracted from the whole dataset without validation on an independent sample. In [40] the author has addressed this issue by introducing an algorithm that searches for association rules in a training set and validates the mined rules on an independent test set.

This approach is applied on a dataset containing medical records of patients with heart disease and results show that this framework significantly reduces the number of irrelevant and redundant rules and produces medically significant rules.

In this study we focus on the use of AR mining for improving clinical text retrieval. The main challenge in using AR in IR is to select the best rules with respect to the query. To address this issue we first extract a list of rules by applying a standard AR mining algorithm then we select the consequent of those rules that contain the query terms in their antecedents. The selected consequents are then used for query expansion and re-ranking to improve the clinical retrieval effectiveness.

2.3. Information Retrieval by Semantic Similarity

Semantic similarity refers to calculating the similarity between two concepts typically by considering their position and their relationship in specific ontology. In [21], authors proposed a novel retrieval model called Semantic Similarity Retrieval Model (SSRM) based on semantic similarity, where queries are expanded with semantically similar terms according to specific threshold. Their experimental results demonstrate performance improvement over traditional information retrieval models, namely Vector Space Model (VSM).

In [16], authors proposed a method for computing semantic similarity that improves the retrieval performance of IR systems in terms of recall and precision. In this method, initial search results are filtered out according to their relevance degree calculated by use

of semantic similarity. Therefore, documents containing concepts that have semantic relationship with the initial keywords receive higher relevance degree. Experimental results demonstrate the effectiveness of this approach.

This study presents a novel medical query context modeling. The query context contains the most related concepts to the original query concepts where each concept is weighted according to its relatedness to the original query concepts. We propose the use of semantic relatedness measure to select the most related concepts to the original query concepts and to weight the related concepts semantically in relation to the original query. The concepts are initially selected using either AR mining or query domain ontology. However they are filtered based on their relatedness score in relation with the query.

This model attempts to fulfill the objective of filling semantic gap among queries and documents. The query context attempts to enhance the classical IR model and to calculate and assign a new score to documents by considering semantics.

2.4. Medical TREC Records Track Participations

Many participants in TREC Medical Record challenge projected different techniques to improve their system's retrieval task.

In [12] author's participation relies on query expansion technique using Rocchio's algorithm coupled with gender and age filtering and semantic query expansion using disease synonyms. The Rocchio [49] relevance feedback algorithm models a way of

incorporating relevance feedback information into the vector space model. Gender and age filtering is based on filtering the results according to gender and age constraints identified in the query keywords. The semantic query expansion technique focuses on adding synonyms of medical disease keywords crawled from Polysearch¹ to the queries. The approach did not have a significant impact on the retrieval performance and the results illustrates that query expansion alone has a slightly positive impact on the performance while the filters have a negative impact on the performance. The lack of significant improvement on the overall retrieval results show that their approach has limited semantic capabilities where a semantic representation of disease and treatment needs further improvements.

In [8] authors participation relies on query expansion via multiple lexicons including ICD-9 (International Classification of Diseases) and Polysearch along with domain knowledge rules. Performance measures imply limited impact on retrieval performance, possibly due to limited semantic capabilities of expansion terms, however results foster the effectiveness of query expansion using combination of multiple external resources to the ad hoc task of medical records track.

In [13] authors' participation relies on applying part-of-speech tagging and UMLS concept extraction at the sentence level using bi- directional greedy dictionary matching for noun phrases and query expansion by inclusion of all concepts appearing below the concept of interest in the UMLS hierarchy, capped at the maximum of 100 concepts.

¹ <http://wishart.biology.ualberta.ca/polysearch/include/diseaseIDlist.txt>

Performance measures suggest that their approach was not promising, possibly due to appending too specific concepts to the original query.

3. Semantic Information Retrieval

One of the characteristics of semantic information retrieval is representing queries and documents via concepts rather than simple keywords. Conceptual representation of documents and queries is capable of addressing specific challenges in clinical IR such as terminology and vocabulary mismatch. Vocabulary mismatch arises from describing and presenting the same concept using different terms and keywords; this may be the result of numerous choices of synonyms or simply different background and educational level of people. For example gastroesophageal reflux disease, gastro-osophageal reflux disease, gastric reflux disease, acid reflux disease, GORD and GERD are all different variations, synonyms or acronyms of the same concept.

To represent documents and queries using set of concepts we need proper medical domain knowledge. Several different domain models are developed to enhance biomedicine and health related computer applications including UMLS. For the purpose of this study we use UMLS to conceptually represent our queries and documents. This choice relies on the very broad coverage, grouping of concepts and hierarchical structure of UMLS. Section 3.1 of this chapter introduces UMLS and it's three different categories.

Section 3.2 introduces MetaMap [5] and describes the mapping procedure and the conceptual indexing of documents and queries.

Besides vocabulary and terminology mismatch issue, in clinical IR it is very crucial to retrieve relevant documents that don't contain query keywords or concepts explicitly but contain keywords and concepts that are semantically related to the query. For example a relevant document may not contain the specific disease indicated in the query but may contain lab results, symptoms or other diseases that result in that disease or are closely related to the disease. Therefore it is valuable to determine those concepts that have a relationship with the original query concepts. In section 3.3 of this chapter we address this issue using our novel approach of modeling query domain ontologies. Section 3.4 describes how to extract representative concepts to information need using query domain ontologies.

3.1. Unified Medical Language System (UMLS)

UMLS is a set of health and biomedical dictionaries, standards and software tools that can facilitate the development of biomedical and health related applications such as electronic health records, classification tools and language translators. Typically UMLS is distributed as knowledge sources or databases and associated software tools or programs. Associated software tools can aid developers to customize the knowledge sources according to their specific purposes. The Metathesaurus, the Semantic Network and the SPECIALIST Lexicon are three different categories of UMLS knowledge sources.

3.1.1. Metathesaurus

The Metathesaurus is a large database containing over one million biomedical and health related concepts which are retrieved from several thesauri, indexed biomedical literature, controlled vocabularies, code sets and classification systems. It contains more than 150 source vocabularies including MeSH, RxNorm CT, SNOMED CT, ICD-9-CM, ICD-10-CM, CPT and LOINC. The Metathesaurus is classified by concept or meaning and assigns a unique concept identifier (CUI) to each concept. It connects each concept to its alternative names and views in other source vocabularies; it also determines appropriate relationships among concepts. The Metathesaurus contains synonymous and non-synonymous relationships between concepts from the same source vocabulary as well as concepts in different vocabularies. In general, the relationships connect closely related concepts, such as those that share some common property or are related by definition. The relationships can be of different types including SY (synonym), RB (broader relationship), RN (narrower relationship), SIB (sibling), CHD (has child relationship), PAR (has parent relationship), RU (related, unspecified) RO (has relationship other than synonymous, narrower, or broader) and etc. All the concepts in Metathesaurus are labeled to at least one semantic type from Semantic Network (<http://ncbi.nlm.nih.gov>).

3.1.2. Semantic Network

Semantic Network categorizes all the concepts in the Metathesaurus by semantic types. It also provides a set of valuable and significant relationships among concepts, namely

semantic relations. The major semantic types are organisms, anatomical structures, biologic function, chemicals, physical objects, events and concepts or ideas. The major semantic relation that represents relationships among semantic types is “IS-A” link. This semantic relation constructs a hierarchy of semantic types from the most general to the most specific semantic type. The major non-hierarchical relations are "physically related to", "spatially related to", "temporally related to", "functionally related to", and "conceptually related to". The Semantic Network consists of 133 semantic types and 54 semantic relations. The network is presented as a tree structure where semantic types are the nodes and semantic relations are the links among the nodes.

3.1.3. SPECIALIST Lexicon

The SPECIALIST lexicon is a general English lexicon that includes a broad range of medical and biomedical terms as well as commonly used English words. It is developed to provide lexical information (syntactic, morphological, and orthographic information for each word) that is needed for Natural Language Processing (NLP) tasks. It also provides NLP tools that are developed to facilitate users in managing lexical variations in medical and biomedical text, which can be used for indexing or NLP applications.

3.2. Conceptual Representation

One of the main objectives in clinical IR is to fill the semantic gap among the queries and documents and go beyond keywords matching. One of the characteristics of semantic

search is conceptual representation rather than simple keywords to enhance the representation of queries and documents. To achieve this, we map queries and documents to their correspondent medical concept in UMLS using MetaMap. This section introduces MetaMap and describes the conceptual indexing of our dataset.

3.2.1. MetaMap

MetaMap [5] is a program developed by National Library of Medicine (NLM), which maps biomedical texts to the UMLS Metathesaurus. MetaMap is one of the foundations of NLM's Medical Text Indexer (MTI), which facilitates both semi-automatic and fully automatic indexing of biomedical texts. It locates all the UMLS concepts associated with terms in biomedical texts using knowledge intensive method based on symbolic, natural language processing and computational linguistics techniques.

3.2.2. Conceptual Indexing

For the purpose of this study the patients record set and the query set are semantically indexed using MetaMap and are converted from the original term-based representation into medical concepts defined by UMLS.

MetaMap breaks down biomedical texts into sentences, phrases, lexical elements and tokens and candidate UMLS concepts are generated based on these modules. The mapping process consists of lexical/syntactic analysis, variant generation, candidate identification, mapping construction and word-sense disambiguation (WSD).

The lexical/syntactic analysis of biomedical texts consists of tokenization, sentence boundary determination and acronym/abbreviation identification, part-of-speech tagging, lexical lookup of modules in the UMLS SPECIALIST lexicon and another final syntactic analysis.

Once phrases are found by lexical/syntactic analysis, they are further analyzed by variant generation, candidate generation, mapping construction and WSD. Variants of each phrase are generated by variant generation. Candidate identification is responsible for evaluating matched concepts. Mapping construction evaluates all candidates and generates final results according to highest ranked candidates. WSD is the process of identifying sense of terms by their position in the sentence and their surrounding text.

Evaluation is based on combination of four measures, which are: centrality, variation, coverage and cohesiveness. Centrality is a Boolean value; it is 1 if the head input text is involved in any of the candidates and 0 otherwise. Variation calculates the difference between the candidate mappings' words and input text words. Variant types include spelling, inflectional, synonym or acronym and derivational. Coverage value indicates how much of input text is engaged in the mapping and cohesiveness measures how many connected component of input text is involved in the mapping. Each measure is normalized to a value between 0 and 1 in which, coverage and cohesiveness receive twice the weight of variation and centrality. The final score is computed based on average of four measures and is normalized to a value between 0 to 1000 in which, 0 indicates no match and 1000 indicates identical match.

Figure 1 demonstrates a sample of MetaMap indexing output. By default, MetaMap demonstrates only those concepts with highest scores where “meta candidate” represents intermediate results whereas “meta mapping” indicates final results that best match the input text. As Figure 1 demonstrates, “patients” is mapped to UMLS concept “C0030705” of type “patient or disabled group” with rank of 1000, which represents identical match. “Complicated” is mapped to UMLS concept “C0231242” of type “functional concept” with rank of 694 and “GERD” is mapped to UMLS concept “C0017168” of type “disease or syndrome” with rank of 861. “Received” is mapped to UMLS concept “C1514756” of type “qualitative concept” with rank of 1000 and “endoscopy” is mapped to UMLS concepts “C1552424” of type “health care related organization, manufactured object” and “C0014245” of type “diagnostic procedure”, both with rank of 1000.

For the purpose of this study queries and documents are indexed according to the described procedures and are represented as CUI’s as opposed to original term-based representation.

```

Phrase: "Patients"
Meta Candidates (Total=1; Excluded=0; Pruned=0; Remaining=1)
  1000 C0030705:Patients [Patient or Disabled Group]
Meta Mapping (1000):
  1000 C0030705:Patients [Patient or Disabled Group]

Phrase: "with complicated GERD"
Meta Candidates (Total=2; Excluded=0; Pruned=0; Remaining=2)
  861 C0017168:GERD (Gastroesophageal reflux disease) [Disease or Syndrome]
  694 C0231242:Complicated [Functional Concept]
Meta Mapping (888):
  694 C0231242:Complicated [Functional Concept]
  861 C0017168:GERD (Gastroesophageal reflux disease) [Disease or Syndrome]

Phrase: "who"

Phrase: "received"
Meta Candidates (Total=1; Excluded=0; Pruned=0; Remaining=1)
  1000 C1514756:Received (Receive) [Qualitative Concept]
Meta Mapping (1000):
  1000 C1514756:Received (Receive) [Qualitative Concept]

Phrase: "endoscopy"
Meta Candidates (Total=3; Excluded=1; Pruned=0; Remaining=2)
  1000 C0014245:Endoscopy (Endoscopy (procedure)) [Diagnostic Procedure]
  1000 C1552424:Endoscopy (Clinic / Center - Endoscopy) [Health Care Related Organization,Manufactured Object]
  928 E C0442418:Endoscopic (Endoscopic approach - access) [Spatial Concept]
Meta Mapping (1000):
  1000 C1552424:Endoscopy (Clinic / Center - Endoscopy) [Health Care Related Organization,Manufactured Object]
Meta Mapping (1000):
  1000 C0014245:Endoscopy (Endoscopy (procedure)) [Diagnostic Procedure]

```

Figure 1. MetaMap Indexing

3.3. Query Domain Ontology

An ontology models knowledge as a set of concepts that are related via specific relationships. Ontologies are often associated with hierarchies of entities and the subsumption relations, but they are not limited to these forms. A domain ontology or domain-specific ontology represents a specific domain, which denotes part of the world. For example the word “chair” has different meanings. An ontology about the domain of furniture would model the seating and accommodation meaning of the word whereas an

ontology about the domain of meeting would model the position of authority meaning of the word.

Modeling and developing medical domain ontologies is a fundamental framework for representing medical knowledge using a set of concepts and the relationships among them. Medical domain knowledge is developed by several different ontologies including UMLS. The structural framework of UMLS ontologies aids us in categorizing concepts and locating related and similar concepts automatically.

In this study we develop a domain ontology for each query. The main motivation for developing domain ontology for each query is to formally represent each query as a set of concepts and identify and represent the relationships between pair of concepts within query domain. This can help us in the retrieval process of relevant documents by capturing not only original query concepts but also concepts that are highly related to the original concepts. Assuming we have a disease in our original query, query domain ontology can help us to locate symptoms of this particular disease, diseases that are associated with this disease as well as related treatments.

To develop query domain ontologies, first we map query keywords to UMLS concepts using MetaMap and second we extract all the concepts that have semantic relationship with the query concept(s) in UMLS Metathesaurus ontologies. The Metathesaurus contains synonymous and non-synonymous relationships between concepts from the same source vocabulary as well as concepts in different vocabularies. In general, the relationships connect closely related concepts, such as those that share some common

property or are related by definition. Semantic relationships can be of different types including SY (synonym), RB (broader relationship), RN (narrower relationship), SIB (sibling), RO (has relationship other than synonymous, narrower, or broader) and etc.

3.4. Query Concepts Extraction using Query Domain Ontology

The main motivation for including synonymous and non-synonymous relationships for modeling query concepts domain ontology is to capture all the concepts closely related to query concept(s). Concepts extraction is capped at second level to avoid concepts that are not in a close relation with query concepts. This ontology is capable of addressing the issue of granularity mismatch by including the parent-child relationships, where the child concept is a subset of the parent concept. This ontology is also capable of tackling the problem of retrieving those relevant documents not containing queries' keywords explicitly but containing concepts semantically related to the queries, since it includes alternative names and views of the same concept as well as concepts that are closely related to the query concept(s).

Figure 2 demonstrates Hearing Impairment domain ontology. To develop this ontology, first query keywords "hearing impairment" are mapped to UMLS concepts using MetaMap. MetaMap suggests "C1384666" as identical match to "hearing impairment". Second, all the concepts that have semantic relationship with "hearing impairment" are extracted using UMLS Metathesaurus. As Figure 2 represents "hearing difficulty" has a relationship of type "RO", "complete hearing loss" has a relationship of type "RO",

“deafness” has a relationship of type “RO”, “sensation of blocked ears” has a relationship of type “RO”, “sensorineural hearing loss” has a relationship of type “RB”, “partially hearing impairment” has a relationship of type “RO”, “hearing problem” has a relationship of type “RN” and “hearing disability” has a relationship of type “RO” with “hearing impairment”. At the second level “hearing difficulty” has a relationship of type “RO” with “encounter due to problems with hearing”. “Complete hearing loss” has relationships of type “RO” with both concepts “deafness symptom” and “hearing loss, bilateral”. “Deafness” has relationships of type “RO” with both concepts “lip reading” and “hearing aid” and has a relationship of “RB” with “neural hearing loss”. “Sensation of blocked ears” has relationships of type “RO” with “blocked ears” and “ear disease”. “Sensorineural hearing loss” has a relationship of type “RN” with “disease of inner ear” and a relationship of type “RB” with “congenital deafness”. “Partially hearing impairment” has a relationship of type “RO” with “partially impaired persons”. “Hearing problem” has relationships of type “RB” with both concepts “middle ear deafness” and “buzzing in ear”.

As Figure 2 suggests, medical query domain ontology is capable of expressing the close relationship of diseases (Hearing Impairment and Hearing Loss), disease and symptom (Hearing Impairment and Buzzing in Ear) and disease and treatment (Hearing Impairment and Hearing Aid). Therefore, according to this ontology if a document doesn't contain hearing impairment concept but contains the associated treatment hearing aid, our proposed model will retrieve it since it contains a concept that is in hearing impairment domain ontology. However related concepts receive a different weight from the original

query concept(s). Concepts extracted from query domain ontology are later used for query reformulation and query context modeling, which is discussed in chapter 5.2.

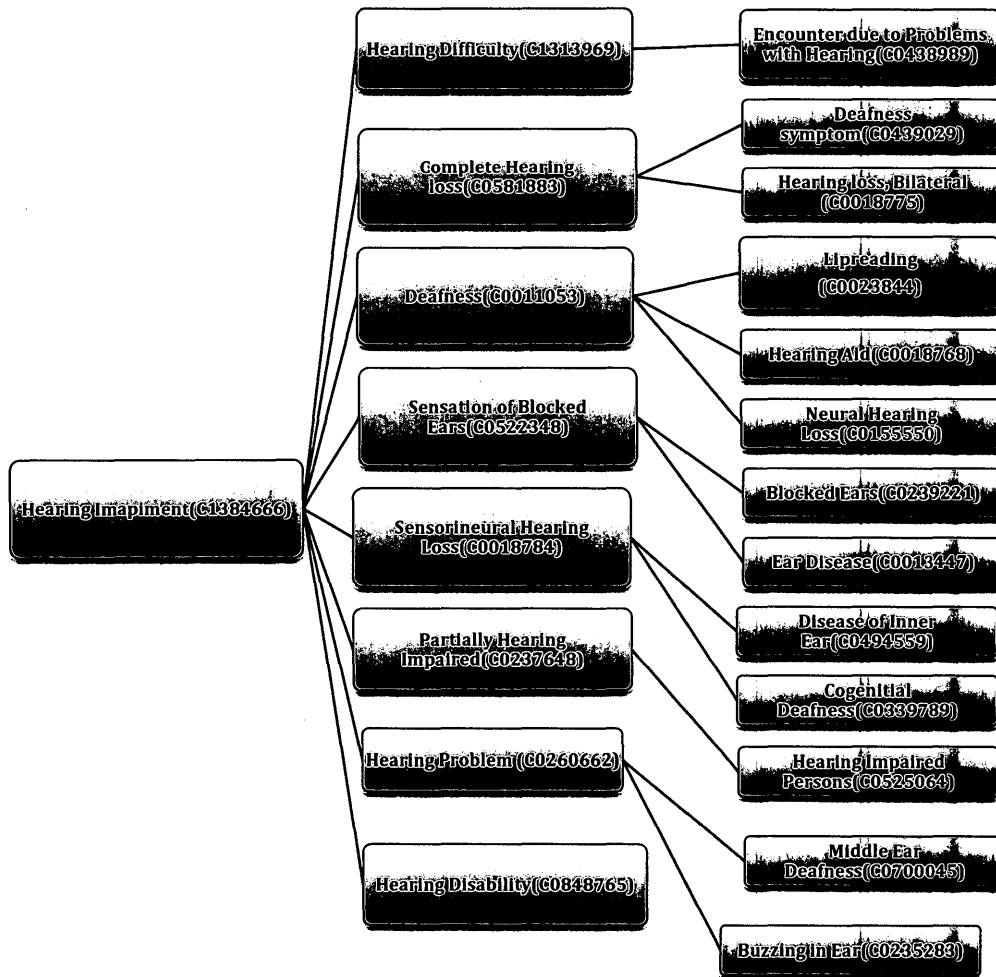


Figure 2. Hearing Loss domain ontology

4. Association Rule Mining in Clinical Retrieval

Association rule mining has been widely used on electronic medical records for discovering hidden patterns. It has also been used for improving information retrieval performance via query expansion. The usefulness of association rule mining is limited by the huge amount and the low quality of generated rules. Many of the generated rules are redundant, trivial or semantically wrong. The main challenging issue in using AR in IR is to select the best rules with respect to the query. To address the presented issues, in this study we propose the use of an AR mining algorithm namely TopKRules algorithm to extract rules that can represent hidden knowledge in patients records. Then, we select the consequents of those rules that cover the query concepts in their antecedents.

In section 4.1 of this chapter we introduce the original definition of AR mining. Section 4.2 discusses the use of AR mining in improving clinical IR and presents our proposed approach of using AR for retrieval effectiveness. Section 4.3 introduces the algorithm we use for discovering hidden patterns in patient records and section 4.4 describes how to extract representative concepts to information need using AR mining.

4.1. Association Rule Mining

One of the most important data mining tasks is association rule (AR) mining, which

attempts to discover interesting associations and relations among data items in large databases. AR mining was first introduced by [1] and its objective is to find all the co-occurrence relationships among data items under the form of implications if X then Y , denoted as $X \rightarrow Y$, where x and y are named antecedent and consequent respectively.

Based on the original definition by [1], the problem of association rule mining is stated as follows. Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of items, where an itemset is a set of items $X \subseteq I$ and $D = \{t_1, t_2, \dots, t_m\}$ be a set of transactions called transaction database, where each transaction has a unique ID and is a subset of I ($t_m \subseteq I$) and consists of items purchased by a customer in a visit. An association rule is denoted as $X \rightarrow Y$ which implies a relationship between two itemsets where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The support of an itemset X is denoted as $sup(X)$ and is defined as the number of transactions that contain itemset X . The support of a rule $X \rightarrow Y$ is defined as $sup(X \rightarrow Y) = \frac{sup(X \cup Y)}{|D|}$. The confidence of a rule $X \rightarrow Y$ is defined as $conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$. The problem of AR mining is to discover all association rules in a transaction database having a support greater or equal to a user specified threshold $minsup$ and a confidence no less than a user specified threshold $minconf$.

Standard association rule mining algorithms have two phases; first phase discovers all frequent itemsets in the database satisfying user specified minimum support, it starts with generating frequent one-itemsets and proceeds to two-itemsets and so on until there are no more frequent itemsets. Second phase generates all association rules using frequent itemsets discovered in previous step. All the generated rules have support and confidence

greater or equal to the user specified threshold.

For example, Figure 3 demonstrates a transaction database and some of the association rules discovered for $minsup=0.4$ and $minconf=0.5$. Step 1 is to generate frequent itemsets in the database whose support $\geq minsup$. For instance, the itemset {Milk, Diaper, Beer} has a support of 0.4 since it occurs in 40% of all transactions (2 out of 5 transactions, transaction 4 and 5). Step 2 is to generate rules from frequent itemsets discovered from step 1 whose confidence $\geq minconf$. For instance, confidence of the rule {Milk, Diaper} \rightarrow {Beer} is $0.4/0.6 = 0.67$, which means that for 67% of the transactions containing milk and diaper the generated rule is correct.

4.2. Association Rule Mining in Information Retrieval

Expanding queries with synonyms, acronyms and associated or related terms aids us in capturing relevant documents and improving performance of retrieval systems. Relevance feedback, pseudo-relevance feedback [9] and using external resources are methods for expanding and reformulating queries. In relevance feedback, users identify relevant documents and related terms are extracted from identified documents. In pseudo-relevance feedback related terms are extracted from top k retrieved documents. The other choice is to extract related terms using external resources such as thesaurus.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

RID	Rules	Support	Confidence
1	{Milk, Diaper} → {Beer}	0.4	0.67
2	{Milk, Beer} → {Diaper}	0.4	1.0
3	{Diaper, Beer} → {Milk}	0.4	0.67
4	{Beer} → {Milk, Diaper}	0.4	0.67
5	{Diaper} → {Milk, Beer}	0.4	0.5
6	{Milk} → {Diaper, Beer}	0.4	0.5

Figure 3. (a) A transaction database and (b) some discovered association rules

In this study, we focus on the use of AR mining to address query reformulation problem and improving clinical information retrieval. In the biomedical field, association rule mining presents a promising technique for finding hidden patterns in a medical dataset and for improving the information retrieval performance via query expansion. The use of AR in IR is challenging since many rules are trivial, redundant, semantically wrong and conflict with common sense or basic domain knowledge, or already known by end-users. The main challenging issue in using AR in IR is to select the best rules with respect to the query. Association rules are selected based on interestingness measures. Interestingness measures select and rank discovered rules according to user's interest. Current interestingness measures consider frequency of items and time and space cost but they ignore semantics, which is the main focus in this study.

In most of previous work, AR mining is applied to address the query reformulation

problem where AR algorithms are applied to documents to discover rules based on specified thresholds and the consequent of the rules that contain the query terms in the antecedents are added to the original query. However these studies disregard semantics for selecting the relevant rules, which is more necessary in medical domain.

For the purpose of this study, we use concept-based representation of documents. To apply AR mining to the retrieval process of clinical documents, we assume that each document is a transaction while each concept inside the document is an item or a product bought by a customer.

To address the query reformulation problem using AR mining, first we retrieve n initial results based on a basic term-based retrieval model. In the second step we extract a list of rules by applying TopKRules algorithm on the n initial results then we select the consequent of those rules that cover the query concepts in their antecedents. The extracted concepts are then weighted according to their semantic similarity/relatedness to the query concepts, which is discussed in chapter 5.

4.3. TopKRules Algorithm

To extract list of rules we apply TopKRules AR mining algorithm to concept-based representation of documents. This choice relies on the fact that depending on user specified thresholds, association rule mining algorithms have disadvantages of being very slow, consuming huge memory and producing huge or insignificant amount of information. TopKRules algorithm is proposed by [14] to address these issues. This

algorithm relies on rule expansions and incorporates numerous optimizations, which improves its performance and scalability.

The inputs to TopKRules algorithm are a transaction database, *minconf* and a number *k* specified by the user, which corresponds to the number of rules to be returned. The algorithm starts with assigning 0 to internal *minsup* variable then it starts searching for rules. Once a rule is found it is added to a list of rules *L* sorted by their support. When *k* valid rules are found, the internal *minsup* is raised to lowest support of association rules in the list. New generated rules based on new *minsup* are added to the list and rules not satisfying the new *minsup* are removed from the list. This process continues until no more rules are found or in other words top *k* rules are found.

This algorithm does not rely on the typical two-phase approach of generating frequent itemsets and generating association rules based on the frequent itemsets, since it is not efficient. The algorithm relies on rule expansion approach. It starts with generating rules with single antecedent item and single consequent item, and then it scans the transaction database in order to find single items that can be appended to left or right side of the association rule. As a result, adding single items to antecedent or consequent expands the rules recursively.

Figure 4 represents the procedure of TopKRules algorithm. First, it starts with scanning the database to calculate *tids* for each single item in the database (*line 1*). *Tids* or sets of transaction ids is defined as $tids(X) = \{t | t \in D \wedge X \subseteq t\}$ where *X* is an itemset, *D* is a database and *t* is a transaction. Then the algorithm generates all the valid rules with

single item in antecedent and single item in consequent (*line 2*). Support of each rule is calculated by dividing $|tids(i \rightarrow j)|$ by $|D|$ and dividing $|tids(j \rightarrow i)|$ by $|D|$ (*line 3 and 4*). Confidence of each rule is calculated by dividing $|tids(i \rightarrow j)|$ by $|tids(i)|$ and dividing $|tids(j \rightarrow i)|$ by $|tids(j)|$ (*line 5 and 6*). Then each valid rule is saved in the list L , which contains the current top k rules (*line 7 to 9*). Next, each rule is added to the set R and a flag named *expandLR* is set to true where R is a set that stores all the rules that can be expanded and *expandLR* denotes the rules should be expanded in right side and/or left side (*line 10 to 12*). Then, rules having the highest support are selected and expanded recursively, where EXPAND-L and EXPAND-R indicate expanding left side of the rule and right side of the rule respectively. This loop stops when there is no more rules in the set R with a support greater than *minsup* (*line 15 to 23*).

For the purpose of this study, we apply this algorithm to obtain top k associations where we assume each document is a transaction and concepts in each document are items.

4.4. Query Concepts Extraction using Association Rule Mining

Query concepts extraction consists of extracting consequent of those rules where their antecedents contain query concepts. To achieve this, we search for the original query concepts in the antecedents of the rules obtained from TopKRules algorithm to find the associated concepts in the consequent of the rules. These concepts are then used for query reformulation and query context modeling, which is discussed in chapter 5.2.

TOPKRULES($D, k, minconf$) $R := \emptyset. L := \emptyset. minsup := 0.$

1. Scan the database D once to record the tidset of each item.
2. FOR each pairs of items i, j such that $|tids(i)| \times |D| \geq minsup$ and $|tids(j)| \times |D| \geq minsup$
3. $sup(\{i\} \rightarrow \{j\}) := |tids(i) \cap tids(j)| / |D|.$
4. $sup(\{j\} \rightarrow \{i\}) := |tids(i) \cap tids(j)| / |D|.$
5. $conf(\{i\} \rightarrow \{j\}) := |tids(i) \cap tids(j)| / |tids(i)|.$
6. $conf(\{j\} \rightarrow \{i\}) := |tids(i) \cap tids(j)| / |tids(j)|.$
7. IF $sup(\{i\} \rightarrow \{j\}) \geq minsup$ THEN
8. IF $conf(\{i\} \rightarrow \{j\}) \geq minsup$ THEN SAVE $(\{i\} \rightarrow \{j\}, L, k, minsup).$
9. IF $conf(\{j\} \rightarrow \{i\}) \geq minsup$ THEN SAVE $(\{j\} \rightarrow \{i\}, L, k, minsup).$
10. Set flag expandLR of $\{i\} \rightarrow \{j\}$ to true.
11. Set flag expandLR of $\{j\} \rightarrow \{i\}$ to true.
12. $R := R \cup \{i\} \rightarrow \{j\}, \{j\} \rightarrow \{i\}.$
13. END IF
14. END FOR
15. WHILE $\exists r \in R$ AND $sup(r) \geq minsup$ DO
16. Select the rule $rule$ having the highest support in R
17. IF $rule.expandLR = true$ THEN
18. EXPAND-L($rule, L, R, k, minsup, minconf$).
19. EXPAND-R($rule, L, R, k, minsup, minconf$).
20. ELSE EXPAND-R($rule, L, R, k, minsup, minconf$).
21. REMOVE $rule$ from $R.$
22. REMOVE from R all rules $r \in R | sup(r) < minsup.$
23. END WHILE

Figure 4. The TopKRules Algorithm

Figure 5 represents the procedure of query concepts extraction where L is the list of rules generated from TopKRules AR mining algorithm, Q_c is the list of concepts in the original query and P is the list of permutations of query concepts. P or the list of permutations contains all of the ordered combinations of query concepts. For example, the query

$q = c_1c_2$ has two ordered combination of c_1c_2 and c_2c_1 . Each item in the permutations list is compared to *LHS* of each rule in L , where *LHS* indicates left hand side of the rule. If *LHS* of a rule is identical to an item in P , *RHS* (right hand side of the rule) will be saved. Output of this procedure is a list of consequent concepts C .

INPUT: List of rules generated from TOPKRules L , List of original query concepts Q_c , List of permutations of query concepts P

OUTPUT: List of consequent concepts C

1. **FOR** each item c in P
 2. **FOR** each item m in L
 3. **IF** $LHS(L_m) = P_c$ **THEN**
 4. **SAVE** $RHS((L_n), C)$
 5. **END FOR**
 6. **END FOR**
-

Figure 5. Query Concepts Extraction

5. Information Retrieval by Semantic Similarity and Semantic Relatedness

Human Judgments regarding relatedness or similarity of pair of concepts can help us improve the performance of information retrieval systems. These kinds of judgments can improve the retrieval performance by retrieving documents that don't contain the query concepts explicitly but contain concepts that are semantically related or similar to the query concepts. Therefore, semantic similarity/relatedness measures that can imitate human judgments are valuable for us and can aid us in enhancing retrieval performance. Section 5.1 of this chapter describes some of the proposed semantic similarity and semantic relatedness measures along with their advantages and disadvantages.

In this chapter we present our novel query context modeling approach. The query context consists of the most related concepts to the query concepts. These concepts are weighted according to their relatedness to the query concepts. The query context attempts to fulfill the objective of filling the semantic gap between queries and documents. Section 5.2 describes our proposed approach of semantic query context modeling. The query context is then exploited in query expansion and patients' records re-ranking for improving

clinical retrieval performance. Section 5.3 and 5.4 explain semantic query expansion and semantic re-ranking respectively.

5.1. Semantic Similarity and Semantic Relatedness Measures

Semantic relatedness and semantic similarity are useful measures for effective natural language processing, artificial intelligence and information retrieval in medical domain [33]. These measures try to mimic human judgments of relatedness and similarity among concepts. They take a pair of concepts as input and return a numeric value that represents how similar or related they are. Several different similarity/relatedness measures have been proposed to automatically measure similarity/relatedness. This section describes some of the proposed approaches.

5.1.1. Semantic Similarity

Semantic similarity can be described as a special case of semantic relatedness. Semantically similar concepts are related on the basis of their likeness, shape or form [43]. Semantic similarity measures determine how similar two concepts are by calculating how close they are in semantic network hierarchies. Some of the proposed similarity measures include path lengths [46], Wu and Palmer [56], Leacock and Chodorow [29] and Hirst and St-Onge [20].

5.1.1.1. Path Lengths

This measure is based on path lengths between concepts in the MeSH ontology. To calculate the similarity of pair of concepts, this measure counts the edges between two concepts. Although this measure is simple and it showed reliable results [43] but it requires a reliable and comprehensive ontology. Another disadvantage of this approach is that it only includes relations of type “IS-A” that connects general concepts to specific concepts and disregards other forms of relations among concepts.

5.1.1.2. Wu and Palmer

This semantic similarity measure calculates the similarity of two concepts by taking into account depth of two concepts in UMLS semantic network as well as Least Common Subsumer (LCS). LCS refers to the deepest shared ancestors of two concepts in a hierarchy. Equation 1 defines Wu and Palmer semantic similarity measure, where c_1, c_2 is the pair of concepts of interest and $0 < score \leq 1$. The score is 1 when two concepts are the same. Disadvantage of this approach is that it only includes relations of type “IS-A”.

$$Score = \frac{2 \cdot depth(LCS)}{depth(c_1) + depth(c_2)}$$

Equation 1

5.1.1.3. Leacock and Chodorow

Leacock and Chodorow's semantic similarity measure is based on finding the shortest path between two concepts. Equation 2 defines Leacock and Chodorow's similarity measure where $length$ is the shortest path between two concepts and D represents the maximum depth of the ontology. Disadvantage of this approach is that it only includes relations of type "IS-A".

$$score = -\log(\frac{length}{(2 \cdot D)})$$

Equation 2

5.1.1.4. Hirst and St-Onge

This similarity measure is a path-based measure that includes more relation types. This measure calculates similarity of two concepts by determining the nature of the path between them. Paths that are shorter and have less change in direction represent higher similarity whereas paths that are longer and have more change in direction represent lower similarity. However, this measure relies on relations that don't exist in UMLS [43].

5.1.2. Semantic relatedness

Semantic relatedness measures are more general than semantic similarity measures; they are not based on likeness, shape or form of concepts. Semantic relatedness measures determine how related two concepts are by using concepts definitions information,

relations information and co-occurrences information. Some of the proposed similarity measures include Resnik [47], Jiang and Conrath [26], [31] and Gloss Vector [42].

5.1.2.1. Resnik

This relatedness measure is based on information content (IC) and LCS. IC can be considered as a measure of the specificity of a concept [42]. Higher value of IC indicates more specific concepts whereas lower value of IC indicates more general concepts. IC is calculated based on frequency counts of concepts as located in a large corpus of text. Equation 3 defines information content of concept c , where $freq(c)$ is the frequency of concept c and $freq(root)$ is the frequency of the root of the hierarchy.

$$IC(c) = -\log\left(\frac{freq(c)}{freq(root)}\right)$$

Equation 3

As Equation 4 demonstrates Resnik relatedness measure is calculated by finding IC of LCS of pair of concepts (c_1, c_2) .

$$Score = IC(lcs(c_1, c_2))$$

Equation 4

Although this measure has the advantage of using empirical information from corpora but it has the disadvantage of using only “IS_A” type of relation. According to this measure concepts that share same LCS would have identical scores, which is not correct. Another disadvantage of this measure is that this measure only includes IC of LCS of pair of concepts rather than IC of individual concepts.

5.1.2.2.Jiang and Conrath

Jiang and Conrath relatedness measure attempts to address the disadvantage of Resnik measure by including IC of individual concepts in the relatedness measure. Equation 5 defines Jiang and Conrath relatedness measure of pair of concepts (c_1, c_2).

$$Score = IC(c_1) + IC(c_2) - 2 . IC (lcs (c_1, c_2))$$

Equation 5

Although this measure takes IC of individual concepts into consideration and solves one of the disadvantages of Resnik measure but it has the disadvantage of using only “IS_A” type of relation.

5.1.2.3.Gloss Vector

This semantic relatedness measure represents concepts by context vectors or vectors of co- occurrences and measures relatedness of concepts by calculating the cosine of the angle between Gloss Vectors. This measure creates Gloss Vectors for each concept presented in WordNet using second order co- occurrences and structure and content of WordNet and WordNet definitions. Gloss Vector is a second order co-occurrence vector created by using WordNet as the dictionary and the corpus of text. First order co- occurrences refer to terms that occur close to each other in a corpus of text. For instance, police and car are likely first order co-occurrences since they usually occur close to each other [42]. Two terms are second order co-occurrences when they are both first order co- occurrences of a term. For example if car and mechanic are first order co-occurrences

then police and mechanic would be second order co-occurrences [42].

The value of semantic relatedness indicates the relatedness of pair of concepts, the higher the value the higher the semantic relatedness. The semantic relatedness of pair of concepts is between 0 and 1, the value is 1 if two concepts are identical and 0 if they are not related. Equation 6 defines Gloss Vector relatedness measure where \vec{v}_1 and \vec{v}_2 are context vectors of concepts c_1 and c_2 respectively and the score of each pair of concepts is calculated as the cosine of the angle between vectors \vec{v}_1 and \vec{v}_2 .

$$Score = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|}$$

Equation 6

One of the advantages of this measure is its reliance on concepts definitions information rather than semantic network hierarchy; therefore it is able to calculate semantic relatedness of pair of concepts in different UMLS ontologies. All of the presented measures can only measure relatedness of nouns included in WordNet, however Gloss Vector measure is able to calculate relatedness of all parts of speech. Moreover, since this measure relies on concepts definitions information rather than paths information, it is able to calculate relatedness of a pair of concepts with non-hierarchical relations. Due to the presented advantages of Gloss Vector measure, in our experiments we use this measure to calculate the relatedness of pair of concepts.

5.2. Semantic Query Context Modeling

Terminology mismatch and granularity mismatch are the most critical issues in effectiveness of retrieval systems. Terminology mismatch or vocabulary problem as described by [15] consists of synonymy and polysemy. Synonymy problem refers to same terms with different meanings and polysemy problem refers to different terms with the same meaning. Several approaches have been proposed to deal with vocabulary problem such as query expansion, results clustering, relevance feedback and word sense disambiguation [10].

In this study we propose modeling a semantic query context for each query in order to handle the presented issues of IR. A semantic query context consists of the most semantically related concepts to the original query, which are weighted according to their relatedness to the original query concepts. A semantic query context is able to capture more relevant documents. It is also able to fill the semantic gap among queries and documents.

Modeling a semantic query context consists of the following steps: (1) extracting the potentially representative concepts using either AR mining or query domain ontology, (2) weighting the extracted concepts by measuring semantic relatedness of the extracted concepts to the original query concepts and (3) selecting the most representative concepts according to the assigned weights.

5.2.1. Query Concepts Weighting

In section 3.4 and 4.4 we described techniques to extract representative concepts. However, among the extracted concepts many are trivial, redundant or semantically wrong. In order to identify the most relevant concepts we use semantic relatedness measure to discriminate the concepts with respect to the queries. To achieve this, we calculate relatedness of the query concepts and the extracted concepts. Figure 6 presents query concepts weighting procedure where Q_c is the list of original query concepts, C is the list of extracted concepts via query domain ontology or AR mining and T is the weight threshold. Output of this algorithm is the query context CTX_q . For each concept in the original query we calculate semantic relatedness of each of the concepts in the original query Q_c and each of the extracted concepts in C . If the relatedness score is greater than the specified threshold, we will add the extracted concept k and the score of relatedness between c and k to the query context CTX_q (lines 1 to 5).

INPUT: List of original query concepts Q_c , List of extracted concepts C , Weight threshold T

OUTPUT: Query context CTX_q

1. **FOR** each item c in Q_c
 2. **FOR** each item k in C
 3. **IF** *Semantic Relatedness* (c, k) $> T$ **THEN**
 3. **SAVE** ($k, \text{Semantic Relatedness}(c, k), CTX_q$)
 4. **END FOR**
 5. **END FOR**
-

Figure 6. Query concepts weighting

5.2.2. Semantic Query Context

The medical query context reflects the most related concepts to the original query; these concepts could represent symptoms, procedures or correlated diseases that are in relation with the original query concepts. Since in this study we proposed two different approaches for query concepts extraction we define separate form of query context for each approach.

5.2.2.1. Modeling Query Context Using Query Domain Ontology

When we use query domain ontology to extract potential query concepts the query context can be defined by Equation 7. In Equation 7, C_i represents the query concept, C_{im} represents first level concepts and C_{imj} represents second level concepts, where i is the number of query concepts, m is the number of first level concepts and j is a number of second level concepts. rel represents the score calculated by measuring semantic relatedness of pair of concepts using Gloss Vector relatedness measure. $rel(C_i, C_{imj})$ represents semantic relatedness of related concept C_{imj} in relation with the query concept C_i ; which is assigned to each concept as its weight.

$$CTX_q = (c_{i1}, rel(c_i, c_{i1})), \dots (c_{im}, rel(c_i, c_{im})), (c_{i1j}, rel(c_i, c_{i1j})), \dots (c_{imj}, rel(c_i, c_{imj}))$$

$$0 < i, m, j < n$$

Equation 7

5.2.2.2. Modeling Query Context Using AR Mining

When we use AR mining to extract potential query concepts the query context can be defined by Equation 8. In Equation 8, c_i represents the original query concept and c_{in} represents the consequent of a discovered rule with c_i as its antecedent. $rel(c_i, c_{ij})$ represents the relatedness score of the consequent concept c_{ij} of the discovered rule and the query concept c_i using Gloss Vector relatedness measure.

$$CTX_q = (c_{i1}, rel(c_i, c_{i1})), \dots (c_{ij}, rel(c_i, c_{ij}))$$
$$0 < i, j < n$$

Equation 8

5.3. Semantic Query Expansion

One of the most effective techniques to improve the performance of IR systems is expanding the original queries with other terms that can retrieve more relevant documents or can form better queries. Several query expansion approaches have been proposed to find new terms associated with the original query terms. Our proposed query expansion approach relies on semantic query context to produce more useful queries and to retrieve more relevant documents.

Our proposed query expansion technique is performed by adding the query context CTX_q to the original query. The retrieval process is then performed according to reformulated query. In this case we assign the weight 1 to the original query concepts. The rest of the

concepts are weighted according to their semantic relatedness to the original query concepts. The purpose of this process is to capture those relevant documents that don't contain the original query concepts but contain highly related concepts to the original concepts.

5.4. Semantic Re-ranking

A typical IR re-ranking can be defined as a post-processing method that considers the initial ranking of documents and exploits additional information to improve the performance of IR systems. A typical re-ranking problem is as follows. Let $D = \{d_1, d_2, \dots, d_n\}$ be the set of documents to be retrieved and $D_{init} \in D$ where D_{init} denotes a set of top k initial results ranked in decreasing order and retrieved by a standard IR model. In order to improve the performance and accuracy of the standard IR model, we can re-rank the top retrieved documents. The goal of re-ranking models is to improve IR models effectiveness by re-ordering the set of top initial results D_{init} .

Our proposed re-ranking technique is performed by integrating the query context CTX_q into the information retrieval process. In our proposed re-ranking technique, k is set to 1000 and the top 1000 documents denotes D_{init} . D_{init} is retrieved according to the original query concepts. However the top 1000 initial results are re-ranked according to the query context CTX_q .

Equation 9 defines our proposed scoring formula. For each document d_k retrieved, we combine the initial score S_i and the conceptual score S_c . The initial score is the score

assigned by standard retrieval model and according to the original queries. The conceptual score S_c is based on cosine measure between the conceptual representation of d_k and CTX_q , where d_k and CTX_q are characterized by vectors \vec{d}_k and \vec{CTX}_q respectively. The cosine measure gives a useful measure of how similar the document d_k and the query context CTX_q are. We combine S_i and S_c using a tuning parameter α to balance the impact of the original score S_i and the new score S_c .

$$S(d_k) = (1 - \alpha) \cdot S_i(q, d_k) + \alpha \cdot S_c(CTX_q, d_k)$$

$$0 < \alpha < 1$$

Equation 9

Equation 10 defines the conceptual score S_c , that is computed using cosine similarity measure between the initial retrieved results and concepts in the semantic-based query context.

$$S_c(CTX_q, d_k) = \cos(\vec{CTX}_q, \vec{d}_k)$$

Equation 10

6. Experimental Settings

Section 6.1 of this chapter explains our information retrieval environment such as the system we used, pre-processing settings, indexing settings, retrieval settings and retrieval models and post-processing settings. Section 6.2 describes UMLS settings and the software we used to access the UMLS. Section 6.3 describes the dataset that includes patients' records and query set. In this section we also explain Medical TREC's retrieval task, relevance judgments and evaluation measures. Section 6.4 explains our parameter tuning for finding the optimal value for parameters such as weight threshold, confidence level and α . Section 6.5 describes and explains our experimental runs.

6.1. Information Retrieval System Settings

In our search experiments, we compare the standard retrieval performance using only the original queries (ignoring any query context) to the contextual search performed using the query context.

We use Terrier [41] for indexing and retrieval. This choice relies on the fact that Terrier is a highly flexible, efficient, and effective open source search engine and readily

deployable on large-scale collections of documents such as TREC collections. It also implements state-of-the-art indexing and retrieval functionalities (<http://terrier.org/>).

6.1.1. Pre-processing Settings

To pre-process the collection we use Terrier's TermPipeline. TermPipeline can remove all the terms that should not be indexed by removing general stop words. It is also able to transform terms to their roots by stemming. Stemming is beneficial in IR applications since morphological variants of terms have similar semantic interpretation and it is better to transform them to a single base meaning. For the purpose of this study we use TermPipeline to apply porter stemming and we remove general stop words implemented in Terrier. We also added specific stop words that are very frequent in medical domain such as patient, hospital, medical, room, diagnosis etc as suggested by [18].

6.1.2. Indexing Settings

To index our collection we use Terrier. Terrier offers both simple indexing and block indexing. For standard probabilistic weighting model we use simple indexing however for proximity or dependence weighting model we use block indexing.

A prerequisite step for applying dependence models is to record positional information of each term. Terrier saves positional information by indexing using blocks where block is a unit of text in a document and it can be of size 1 or larger. A block of size 1 can determine the exact position of each term and a block size of n determines the position of

every n terms. For the purpose of this study we use block indexing with blocks of size 1 in order to record the exact position of each term.

6.1.3. Retrieval Settings

The standard term-based retrieval is based on the BM25 probabilistic model. In general probabilistic retrieval models rank documents in decreasing order of probability of relevance to the query where this relevancy depends on the query and document representation [50]. BM25 scoring formula, ranks documents according to query keywords appearing in documents regardless of order of the keywords and their relationships. Equation 11 defines BM25 scoring formula where Q is a query containing keywords q_1, \dots, q_n . $f(q_i, D)$ represents frequency of keyword q_i in document D , $|D|$ represents length of document D and $avgdl$ represents average documents length. b and k_1 are free parameters and we use Terrier's default parameters of $b = 0.75$ and $k_1 = 1.2$.

$$Score(D, Q) = \sum_{i=1}^n IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

Equation 11

Equation 12 defines inverse document frequency IDF of the query keyword q_i . N represents total number of documents and $n(q_i)$ represents number of documents that contain the query keyword q_i .

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

Equation 12

To examine the impact of considering the order of the keywords and their relationships or dependencies we repeat the standard term-based retrieval using DFR-based (Divergence from Randomness) dependence model. DFR IR models are based on the idea of "the more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the term t in the document d " [3]. As opposed to BM25 probabilistic model, which assumes keywords are independent, dependence models incorporate keyword dependency in the retrieval framework. DFR-based dependence model offers two possible options for retrieving documents considering their dependency, full dependency and sequential dependency. Full dependency assumes all the query terms are dependent whereas sequential dependency assumes only neighbor query terms are dependent [44].

Equation 13 defines DFR-based dependence model that is used in our experiments. $Score(D, Q)$ represents the score of document D in respect to query Q and $Score(D, t)$ represents the score of query term t in the document D . p denotes a pair of query terms and $Score(D, p)$ represents the score p in the document D . λ_1 and λ_2 are tuning parameters. Q_2 is a set of pair of query terms, for full dependency Q_2 is a set of unordered pair of query terms and for sequential dependency Q_2 is a set of ordered pair of query terms.

$$Score(D, Q) = \lambda_1 \cdot \sum_{t \in Q} Score(D, t) + \lambda_2 \cdot \sum_{p \in Q_2} Score(D, p)$$

Equation 13

The score of pair of query terms in a document is calculated according to Equation 14. p_{p_1} represents the probability that the document D contains the pair of query terms p given number of times. p_{p_2} represents the probability that the pair of query terms p appears in the document D once again after given number of times.

$$Score(D, p) = -\log_2(p_{p_1}) \cdot (1 - p_{p_2})$$

Equation 14

6.1.4. Query Expansion Settings

In our experiments, we also use Rocchio's [49] query expansion technique; which is implemented in Terrier. Equation 15 [11] defines Rocchio's algorithm; which is a relevance feedback technique and is based on vector space model. \vec{Q}_m represents the modified query vector and \vec{Q}_0 represents the original query vector. \vec{D}_j represents related documents vector, \vec{D}_k represents non-related document vector, D_r is a set of related documents and D_{nr} is a set of non-related documents. a is original query weight, b and c are related documents weight and non-related documents weight respectively.

$$\vec{Q}_m = (a \cdot \vec{Q}_0) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\vec{D}_j \in D_r} \vec{D}_j \right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\vec{D}_k \in D_{nr}} \vec{D}_k \right)$$

Equation 15

This method extracts the most informative terms from the top-returned documents from initial search and expands the original query using the extracted terms. In Terrier's implementation of Rocchio's algorithm there are two parameters that can be set. First, the number of terms to expand a query with and second the number of top-ranked documents from which these terms are extracted. In our experiments we use the default value of 10 as the number of terms to expand a query and the default value of 3 as the number of top-ranked documents.

6.2. UMLS Settings

UMLS::Interface and UMLS::Similarity are freely available Perl modules. We use these two modules in our experiments to access the UMLS and to use the functions and the measures that are implemented in these modules.

6.2.1. UMLS::Interface

UMLS::Interface is a Perl interface to the UMLS. The main purpose of UMLS::Interface module is to retrieve information about concepts such as path information or definitional information. The UMLS::Interface module provides several functions that can be used to extract information about UMLS concepts. Functions that are used in this study are:

- `getRelations`: Input to this function is a CUI. This function returns a list of relations of the input concept.

- `getRelated`: Inputs to this function are a CUI and a relation type. This function returns a list of concepts that are related to the input concept through the specified relation.
- `getPreferredTerm`: Input to this function is a CUI. This function returns the preferred term of the input concept.

6.2.2. UMLS::Similarity

In our experiments we use freely available software `UMLS::Similarity`, which is a Perl module that implements several semantic similarity and semantic relatedness measures based on ontologies in the UMLS. This module relies on `UMLS::Interface`, since it needs path information and definitional information to calculate similarity and relatedness of pair of concepts. This module takes two terms or two CUIs as input and calculates the similarity/relatedness between them according to specified measure.

For example Hearing Impairment (C1384666) and Deafness (C0011053) have the relatedness score of 0.9066 using Gloss Vector relatedness measure, which indicates they are highly related. However Hearing Impairment (C1384666) and Gastro Esophageal Reflux disease (C0017168) have the relatedness score of 0.3112 using Gloss Vector relatedness measure, which indicates they are not related.

6.3. Dataset

The corpus that is used to develop and test our approach is provided within the context of

the TREC (Text Retrieval Conference) medical records track challenge. The Text Retrieval Conference is an ongoing series of workshops and its purpose is to support research in different areas of information retrieval techniques and retrieval environments and to provide the essential infrastructure for large-scale text retrieval including large test collections, evaluation software and uniform scoring procedures. This conference helps to accelerate the transmission of new methodologies from research labs to commercial search engines. Most of the technologies implemented in current's commercial search engines are first developed in TREC (www.trec.nist.gov/overview.html).

Previous TREC participants assessed a variety of information retrieval areas such as multimedia retrieval, question answering and cross-language retrieval. In 2011, TREC has released eight different tracks including chemical IR track, entity track, legal track, crowd sourcing track, microblog track, session track, web track and medical records track.

The goal of the medical records track is to promote research on providing content-based access to the text fields of electronic medical records (<http://trec.nist.gov/pubs/call2011.html>). The ability to search through text-based content of EHRs improves the quality of clinical care and accelerates the diagnosis process and enhances effective access to patient's records. TREC medical records track is composed of patient records, query set and relevance judgments. In this section we first describe the test collection and then we clarify the retrieval task.

6.3.1. Patients Records

The patients' record is a set of de-identified EHRs that made available for research purposes through the university of Pittsburg. EHRs are usually coded and organized based on structural fields to allow easier and more effective access. However the majority of the content is embedded in the non-structured text-based field of the records.

Patients' records cover one month of reports from several hospitals. There is a total of 17,264 visits and 93,551 reports in which, each visit contains between 1 to 415 reports. The original corpus is organized by individual reports and there is a many-to-one relationship among reports and a visit, where a visit is an individual patient's single stay at hospital [35] and is associated with a set of patients' records. The University of Pittsburgh provides a table of this mapping and a visit is used as the unit of retrieval in the track. For the purpose of this study we generate visit-based documents where each document is composed by concatenating the patient reports associated with it.

Each document contains the following data elements that can be used for the retrieval task:

- Checksum, which is the unique report id,
- Type, which is general descriptor of the report and is represented by report codes. There are nine types of reports: radiology reports, history and physicals, consultation reports, emergency department reports, progress

notes, discharge summaries, operative reports, surgical pathology reports, and cardiology reports.

- Subtype, which is more precise descriptor of the report,
- Chief complaint, which describes the main symptom, group of symptoms or reasons for which the patient seeks treatment,
- Admit diagnosis, which is medical diagnoses made by physicians as a basis for continuing patient's assessment and is represented as ICD-9 codes,
- Discharge diagnosis, which is medical diagnoses made by physicians that contains information about patient's identified conditions and is represented as ICD-9 codes,
- Year, which represents the year that the report was recorded,
- Download time, which represents the date that the report was downloaded,
- Update time, which represents the date that the report was updated,
- Deid, which is the unique de-identification ID and
- Report text which is in free text form and is the central part of each report describing symptoms, signs, diseases, family history, lab results and so on.

Figure 7 demonstrates a sample of an EHR of type "operative reports".

```

<report>
<checksum>20051127OP-cQsnkGImzZbN-848-71049104</checksum>
<subtype>ORTHO OP</subtype>
<type>OP</type>
<chief_complaint>LFT LEG PAIN</chief_complaint>
<admit_diagnosis>730.27</admit_diagnosis>
<discharge_diagnosis>250.81,707.14,403.91,428.0,711.06,276.7,424.1,416.0,730.27,250.51,362.01,4
14.8,244.0,272.4</discharge_diagnosis>
<year>2007</year>
<download_time>2009-10-05</download_time>
<update_time/>
<deid>v.6.22.08.0</deid>
<report_text> Report de-identified (Safe-harbor compliant) by De-ID v.6.22.08.0  **INSTITUTION
ORTHOPEDIC SURGERY OPERATIVE REPORT  PATIENT NAME:  **NAME[AAA, BBB M]
ACCOUNT #:  **ID-NUM  **ROOM SURGEON:  **NAME[WWW XXX], M.D. ASSISTANT(S):
**NAME[RRR QQ], M.D.  ATTENDING PHYSICIAN:  **NAME[ZZZ M YYY] SURGERY DATE:
**DATE[Nov 27 05] ADMISSION DATE:  **DATE[Nov 22 2007] DISCHARGE DATE:  PROCEDURES:
TITLE OF OPERATION:  IRRIGATION AND DEBRIDEMENT OF LEFT KNEE.  ANESTHESIA:  General.
COMPLICATIONS:  None.  PREOPERATIVE DIAGNOSIS(ES):  SEPTIC ARTHRITIS, LEFT KNEE.
POSTOPERATIVE DIAGNOSIS(ES):  SEPTIC ARTHRITIS, LEFT KNEE.  HISTORY AND
INDICATIONS:  The patient is a **AGE[in 60s]-year-old female with a history of end-stage renal disease and
hemodialysis with vasculopathy who by history, examination, and laboratory studies had a septic arthritis of the
left knee.  Preoperatively, I spoke to the patient at great length.  I spoke to her and her daughter about the risks
and benefits of surgical intervention.  After thorough a discussion about the risks and benefits of surgery, the
patient gave informed consent.
DESCRIPTION OF OPERATION:  The patient was identified as the patient.  She was taken to the operating
room where she was placed supine on a table.  Anesthesia had attempted to place a block; however, this did not
work and therefore she needed to be intubated.  After successful intubation, a nonsterile tourniquet was carefully
placed high in the left thigh.  The left leg was then prepped and draped in the usual sterile fashion while making
sure to isolate the left foot on which she had surgery a few days prior.  The leg was elevated for 120 seconds and
then the tourniquet was inflated.  A small approximately 5 cm parapatellar arthrotomy was performed sharply
with a knife.  This was taken down into the joint sharply.  The patient was awakened from anesthesia.  Earlier the
tourniquet had been deflated prior to closure.  There were no complications during this procedure.
</report_text>
</report>

```

Figure 7. Sample of an EHR

6.3.2. Query Set

Topics are developed by physicians who are students in the Oregon Health & Science University (OHSU) Biomedical Informatics Graduate Program. Their goal was to develop topics where each topic has a reasonable number of correspondent visits. Topics exploit information from report_text element, they are designed in such a way that are not answerable only by the ICD-9 codes in discharge diagnosis and admit diagnosis elements.

The query set contains 34 topics and each topic specifies a particular disease/condition set

and/or a particular treatment/intervention set. For some topics demographic information such as age and gender is specified as well. Several example topics are shown in Figure 8.

102: Patients with complicated GERD who receive endoscopy
106: Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer
112: Female patients with breast cancer with mastectomies during admission
119: Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes
122: Patients who received total parenteral nutrition while in the hospital
134: Patients admitted with chronic seizure disorder to control seizure activity

Figure 8. Example topics from TREC 2011 Medical Records Track

6.3.3. Retrieval Task

The retrieval task of the medical records track is an ad hoc search task that can be used to identify cohorts for comparative effectiveness research (<http://trec.nist.gov/pubs/call2011.html>). According to each topic the designed system should be able to return a list of visits ranked by decreasing likelihood that the patient's record satisfies the topic's conditions.

6.3.4. Relevance Judgments

Relevance judgments are produced by assessors who are similar students that developed the topics. Relevance judgments are binary and reflect whether a visit is relevant or not with respect to the query. Assessors were instructed to judge each visit and decide whether such a visit would be relevant to a specific topic or not. A relevant judgment

means that the visit is a candidate for the topic whereas a non-relevant judgment means that the visit is not a candidate for the topic.

6.3.5. Evaluation Measures

TREC provides standard evaluation tool for each track. Given the results file and the relevance judgments file, the evaluation tool evaluates an ad hoc retrieval run and reports performance measures. Our runs are evaluated using TREC's medical records track official measures: MAP, bpref, R-prec, P@5 and P@10.

- *MAP*: The MAP measure or Mean average precision for a set of queries is defined as the mean of the average precision for each query, where average precision is defined as the average of the precision values at different recall points where a relevant document was retrieved [51]. Equation 16 defines MAP measure where Q is the number of queries.

$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}$$

Equation 16

- *Bpref*: The bpref measure is intended for situations where relevance judgments are identified to be far from complete (<http://trec.nist.gov/pubs/trec16/appendices/measures.pdf>). bpref measures a relation of whether judged relevant documents are retrieved before judged irrelevant documents. Equation 17 defines the bpref measure, where R is the

number of judged relevant documents, r is a judged relevant document, N is the number of judged irrelevant documents and n is a judged irrelevant document.

$$bpref = \frac{1}{R} \sum_r \left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)} \right)$$

Equation 17

- *R-prec*: R-Precision is the precision after R documents have been retrieved, where R is the number of relevant documents for the topic (<http://trec.nist.gov/pubs/trec16/appendices/measures.pdf>). Equation 18 defines precision. Precision evaluates the quality of relevant retrieved documents but de-emphasizes the exact order or rank of the retrieved documents therefore it is useful for TREC collections where there are huge number of relevant documents. R-Precision is calculated by measuring the average of the R-Precisions of all the topics.

$$Precision = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}}$$

Equation 18

- *P@5*: P@5 is the average of precision at 5 documents retrieved. In other words P@5 is defined as the number of relevant retrieved documents out of top 5 retrieved documents.
- *P@10*: P@10 is the average of precision at 10 documents retrieved.

6.4. Parameter Tuning

For query context modeling using query domain ontologies we conduct experiments to find the optimal value for the following parameters:

Alpha(α): To tune the parameter α , we vary it in [0 1] in Equation 9. Figure 9 presents performance measures in different values of α . According to Figure 9 we can confirm that the best value for parameter α is 0.2.

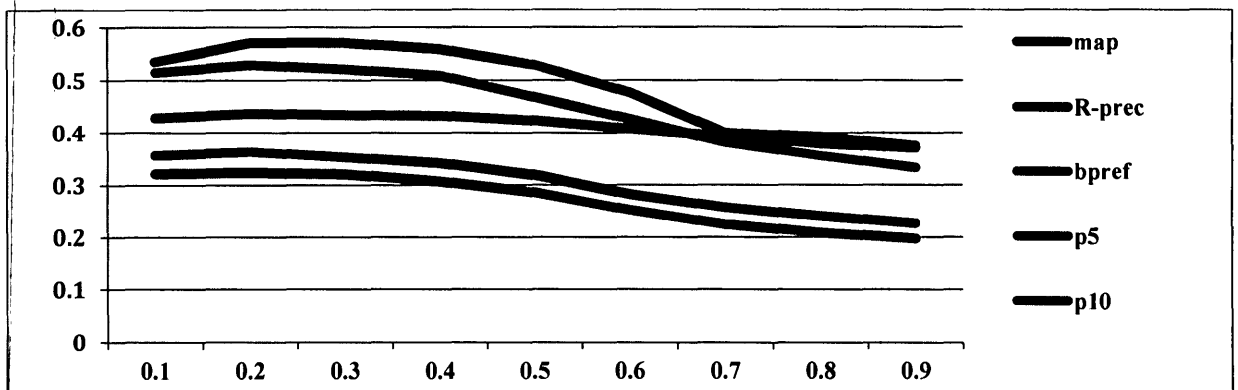


Figure 9. Comparison of parameter α in query domain ontology runs, X-axis indicates parameter α and Y-axis indicates the value of performance measures

Weight threshold: Query context contains concepts with weight or semantic relatedness higher than or equal to a weight threshold. In order to find the optimal threshold for weight restrictions, we conduct different runs to evaluate the impact of this parameter on retrieval performance. Figure 10 presents the impact of modeling query context based on weights equal to or higher than 0.7, 0.75, 0.8, 0.85 and 0.9. For each value we perform conceptual re-ranking of results obtained from query expansion and results confirm that the optimal weight threshold to model query context is 0.9.

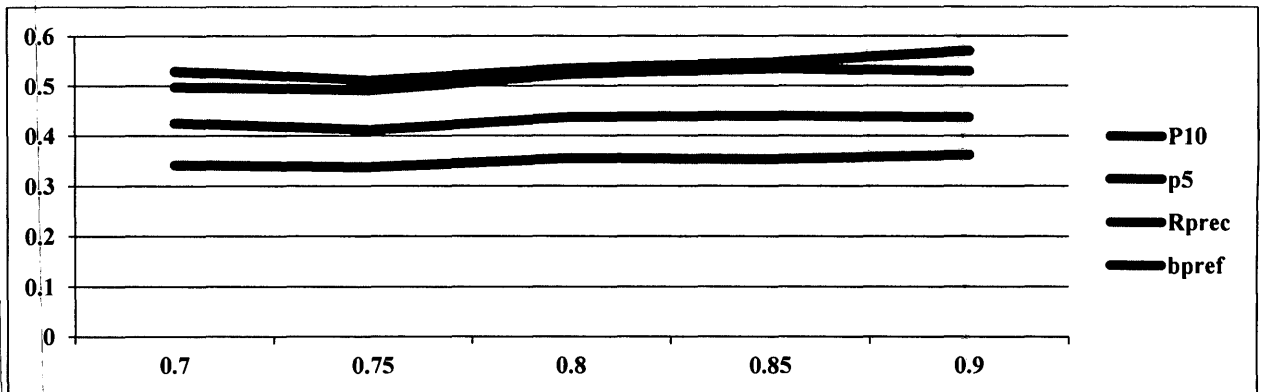


Figure 10. Comparison of weight restrictions, X-axis indicates parameter weight threshold and Y-axis indicates value of performance measures

For query context modeling using association rule mining we conduct experiments to find the optimal value for the following parameters:

Alpha(α): In this experiment, we vary α in [0 1] in Equation 9. For each value we perform concept extraction based on Figure 5 and using top 20 visits. To model the query context we weight the extracted concepts according to Figure 6. Top 10 concepts are used to model the query context. Figure 11 confirms that the optimal value for α is 0.1.

Confidence level: For all the runs *minconf* 0.6 is chosen and this choice relies on the fact that setting higher *minconf* results in missing some interesting rules and setting lower *minconf* results in huge number of rules, which may contain irrelevant and insignificant information.

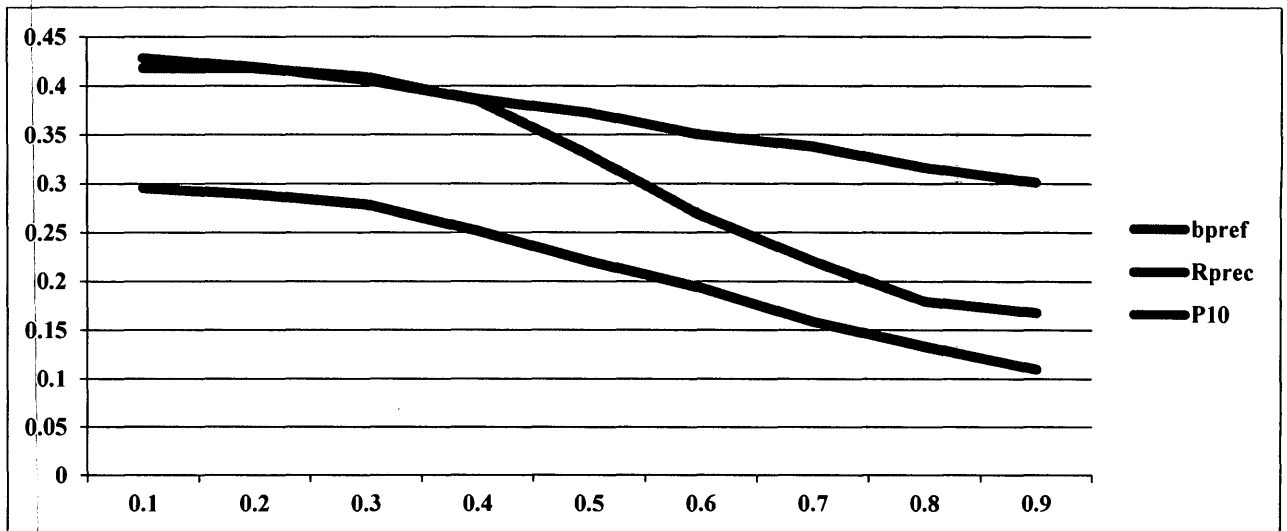


Figure 11 . Comparison of parameter α in association rule mining runs, X-axis indicates parameter α and Y-axis indicates the value of performance measures

Number of documents: In order to find the optimal number of documents used as transactions for AR mining algorithm, we conduct different runs to evaluate the impact of this parameter on retrieval performance. We vary the value for the number of documents in the following set {20,50,100,500,1000}. For each value we perform our contextual ranking using all the extracted concepts based on Figure 5 algorithm. These concepts are then weighted based on Figure 6 algorithm. Re-ranking is based on Equation 9 and $\alpha = 0.1$. As Figure 12 demonstrates the optimal number of documents is top 20 documents per query.

Number of concepts: Figure 13 presents the impact of using top 5, top 10, top 20, top 30 and top 40 concepts on the retrieval performance. All the runs are based on the top 20 documents retrieved from the initial search. The figure reveals that the best performance

is achieved using top 10 concepts; therefore we believe that for the purpose of this study the optimal number of concepts in the query context is 10.

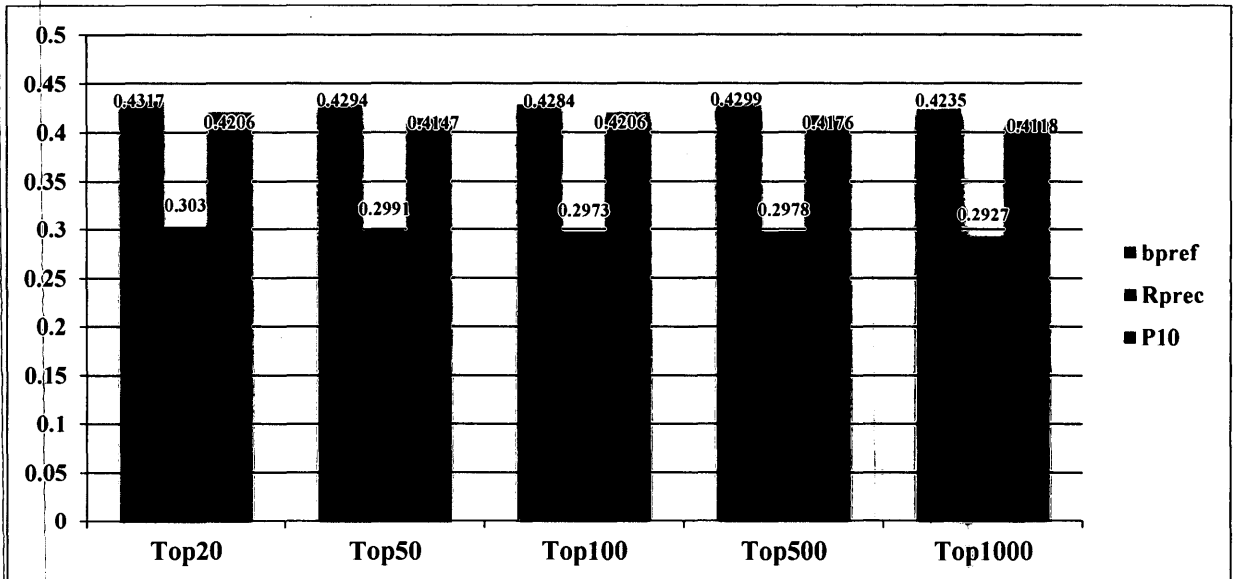


Figure 12 Comparison of performance measures of using top 20, top 50, top 100, top 500 and top 1000 concepts for query context modeling

6.5. Experimental Runs

This section explains and describes a set of runs conducted on the TREC 2011 Medical Records Track. Runs are divided in three sets of baseline runs, query expansion runs and semantic query context runs.

6.5.1. Baseline Runs

- *BL-TermBased-BM25*: This run is based on okapi BM25 probabilistic model where we use Terrier's default parameters $b=0.75$ and $k_1=1.2$. This run is performed using term-based index and the original query set.

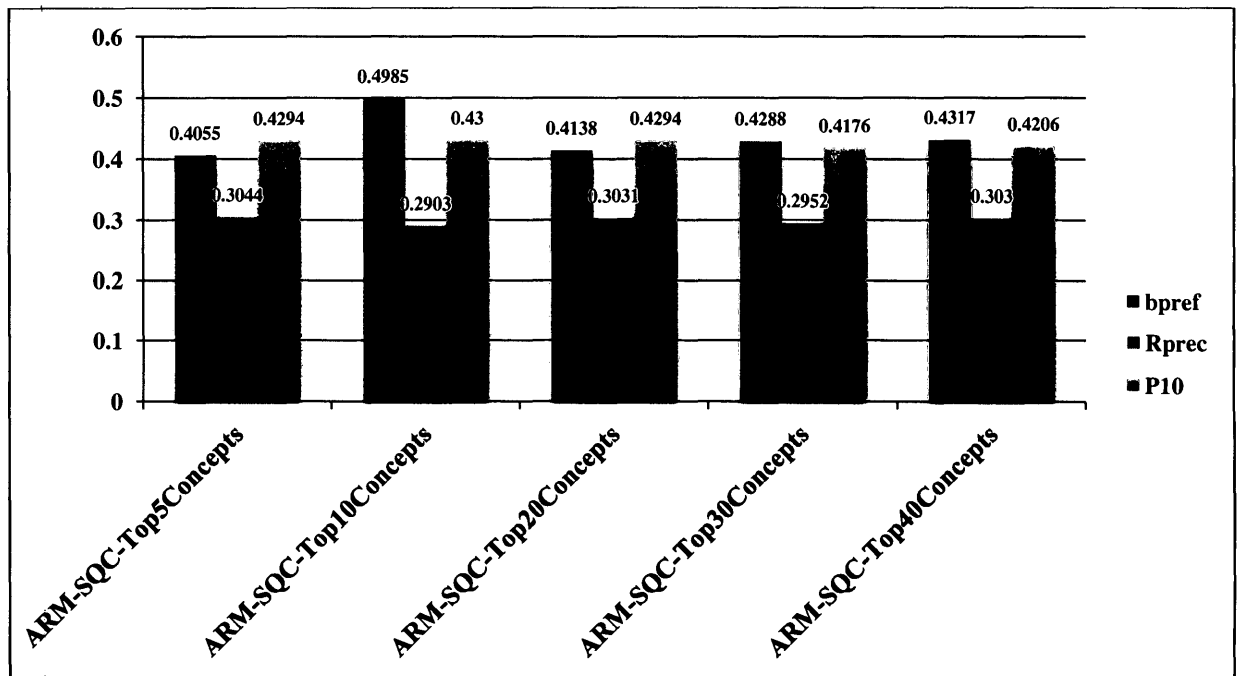


Figure 13 Comparison of performance measures of using top 5, top 10, top 20, top 30 and top 40 concepts for query context modeling

- *BL-ConceptBased-BM25*: This run is based on okapi BM25 probabilistic model where we use Terrier's default parameters of $b=0.75$ and $k_1=1.2$. This run is performed using concept-based index and the original query set.

- *BL-Terms-Concepts-BM25*: This run is based on okapi BM25 probabilistic model where we use Terrier's default parameters of $b=0.75$ and $k_1=1.2$. This run is performed using terms and concepts index and the original query set.
- *BL-TermBased-DFR*: This run is based on DFR sequential dependence model. This run is performed using term-based index and the original query set.

6.5.2. Query Expansion Runs

- *QE-BM25*: This run is based on BM25 probabilistic model where we use Terrier's default parameters of $b=0.75$ and $k_1=1.2$. This run is performed using terms and concepts index and the original query set. Query expansion is performed using Rocchio's relevance feedback algorithm where we use Terrier's default value of 40 as the number of terms to expand a query and the default value of 10 as the number of top-ranked documents from which these terms are extracted.
- *QE-DFR*: This run is based on DFR sequential dependence model. This run is performed using term-based index and the original query set. The reason of using term-based index as opposed to terms and concepts index for this run is that the DFR dependence model develops a framework for term dependencies. Query expansion is performed using Rocchio's relevance feedback algorithm where we use Terrier's default value of 40 as the number of terms to expand a query and the default value of 10 as the number of top-ranked documents.

6.5.3. Semantic Query Context Runs

The following runs are based on our proposed approach of modeling a semantic query context for improving IR systems performance. Runs are divided in two sets: (1) AR mining runs and (2) query domain ontology runs, which represent their method of concepts extraction.

6.5.3.1. Association Rule Mining Runs

In the following runs we extract the representative concepts using AR mining. In order to extract meaningful rules for each query, TopKRules AR mining algorithm is applied on the top retrieved documents from the initial search, where the initial search is *BL-TermBased-BM25* run. The consequents of those rules that cover the query concepts in their antecedents are extracted for modeling the query context.

These runs are performed based on concept-based index. In order to obtain the most precise rules involving only medical related concepts, for the following runs we use only the concepts of five semantic groups. The semantic groups we use for these runs are DISO (Disorders), PROC (Procedures), PHYS (Physiology), CHEM (Chemical and Drugs) and ANAT (Anatomy). Hence we disregard general concepts such as CONC (Concepts and Ideas), GEOG (Geographic Areas) and LIVB (Living Beings). The following runs are conducted in order to evaluate the impact of semantic-based AR mining on the retrieval performance. The following runs are based on the re-ranking formula in Equation 9.

- *ARM-Supp-Top40Concepts*: This run relies on re-ranking the initial search results. In this run we apply TopKRules (k=1000) algorithm on the top 20 documents retrieved from the initial search to extract meaningful rules. The consequents of those rules that cover the query concepts in their antecedents are extracted and ranked according to their support. The query context consists of the top 40 consequent concepts. In this run concepts are weighted according to their support. Re-ranking is based on Equation 9 where $\alpha=0.1$. This value was found to be effective in our experimental settings.
- *ARM-SQC-Top10Concepts*: This run relies on re-ranking the initial search results based on the semantic query context. In this run we apply TopKRules (k=1000) algorithm on the top 20 documents retrieved from the initial search to extract meaningful rules. The consequents of those rules that cover the query concepts in their antecedents are extracted and ranked according to their semantic relatedness to the query concepts. In this run we use Gloss Vector semantic relatedness measure. The query context consists of the top 10 consequent concepts. In this run concepts are weighted according to their semantic relatedness to the query concepts. Re-ranking is based on Equation 9 where $\alpha=0.1$.
- *Naïve-Top10Concepts*: This run relies on re-ranking the initial search results. In this run we extract all the concepts from top 20 documents and rank them according to their semantic relatedness to the query concepts. In this run we use Gloss Vector semantic relatedness measure. The query context consists of the top 10 concepts. In this run concepts are weighted according to their semantic

relatedness to the query concepts. This run is performed to investigate the effectiveness of AR mining concept extraction over naïve method of concept extraction. Re-ranking is based on Equation 9 where $\alpha=0.1$.

6.5.3.2. Query Domain Ontology Runs

In the following runs we extract the representative concepts using query domain ontology. In order to extract the query context concept, we develop a domain ontology for each query. The query domain ontology is modeled and developed from UMLS Metathesaurus. Query domain ontology represents concepts that have a relationship with the query concepts. These concepts are weighted according to their semantic relatedness to the original query concepts. The query context consists of concepts in the query domain ontology with relatedness score above a specific threshold.

- *QE-SQC-DFR*: This run relies on query expansion based on the semantic query context. The query expansion is performed by integrating the semantic query context into DFR-based sequential dependence model. This run is based on weight threshold of 0.9, which was found to be effective in our experimental settings.
- *RR-SQC-DFR*: This run relies on re-ranking the initial search results based on the semantic query context. The initial search is the *BL-TermBased-DFR* run. This run is based on the re-ranking formula in Equation 9. The re-ranking is based on $\alpha=0.2$, which was found to be effective in our experimental settings. This run is based on weight threshold of 0.9.
- *QE-RR-DFR*: This run relies on re-ranking the results obtained from *QE-SQC-*

DFR run. This run is based on the re-ranking formula in Equation 9 where $\alpha=0.2$.

- *QE-RR-DFR-Rocc*: This run relies on re-ranking the results obtained from applying Rocchio's algorithm on *QE-SQC-DFR* run. Re-ranking is based on Equation 9 with $\alpha=0.2$. For Rocchio's query expansion mechanism we use Terrier's default value of 40 as the number of terms to expand a query and the default value of 10 as the number of top-ranked documents from which these terms are extracted.

7. Experimental Results

In this chapter we report the results obtained from the runs described in section 6.5. Our evaluation objective is to assess the effectiveness of our proposed model and examine the extent of its usefulness in improving the performance of clinical IR systems. Results of association rule mining runs are reported in section 7.1 and results of query domain ontology runs are reported in section 7.2.

7.1. Performance of Semantic Query Context Based on AR Mining

This section reports the results obtained from association rule mining runs along with baseline runs and query expansion runs. Our evaluation objectives consist of (1) evaluate the performance of our approach compared to a baseline approach, (2) evaluate the effectiveness of AR mining for query context modeling on the retrieval performance, (3) evaluate the impact of semantic relatedness measures for query context modeling on the retrieval performance, (4) compare our approach to a query expansion technique and a re-ranking using basic AR mining and (5) evaluate the performance of our approach

compared to a well-known query expansion approach.

To achieve our evaluation objectives runs demonstrated in Table 1 are conducted. The *BL-ConceptBased-BM25* run is conducted to evaluate the impact of using concept-based index over term-based index. The *BL-Terms-Concepts-BM25* run is conducted to evaluate the impact of using terms and concepts index over concept-based index and term-based index. The *ARM-Supp-Top40Concepts* run is performed to evaluate the effectiveness of AR mining for query context modeling on the retrieval performance. The *ARM-SQC-Top10Concepts* run is performed to evaluate the effectiveness of semantic AR mining and the impact of semantic relatedness measures for query context modeling on the retrieval performance. The *Naïve-Top10Concepts* run is performed to evaluate the effectiveness of semantic relatedness measures for query context modeling without applying association rule mining. The *QE-BM25* run is conducted to evaluate the performance of our approach compared to a well-known query expansion approach. Table 1 demonstrates the performance of these runs.

7.2. Performance of Semantic Query Context Based on Query Domain Ontology

This section reports the results obtained from query domain ontology runs along with baseline runs and query expansion runs. Our evaluation objectives consist of (1) evaluate the performance of our approach compared to a baseline approach, (2) evaluate the effectiveness of query expansion using semantic query context based on query domain ontology, (3) evaluate the effectiveness of re-ranking using semantic query context based

on query domain ontology, and (4) evaluate the performance of our approach compared to a well-known query expansion approach.

Table 1 Performance evaluation of semantic query context based on association rule mining; results are compared to *BL-TermBased-BM25*.

Runs	MAP	Rprec	Bpref	P@5	P@10
<i>BL-TermBased-BM25</i>	0.2343	0.2551	0.3646	0.4059	0.4000
<i>BL-ConceptBased-BM25</i>	0.1769 (-24%)	0.2255 (-11%)	0.3742 (+2.5%)	0.3471 (-14%)	0.3324 (-17%)
<i>BL-Terms-Concepts-BM25</i>	0.2925 (+25%)	0.3084 (+21%)	0.4359 (+19.5%)	0.4471 (+10%)	0.4118 (+3%)
<i>ARM-Supp-Top40Concepts</i>	0.2647 (+13%)	0.3030 (+19%)	0.4317 (+18.5%)	0.4529 (+11.5)	0.4206 (+5%)
<i>ARM-SQC-Top10Concepts</i>	0.2804* (19.5%)	0.2903 (+14%)	0.4985* (+37%)	0.4600 (+13%)	0.4300 (+7.5%)
<i>Naïve-Top10Concepts</i>	0.2582 (+10%)	0.3044 (+19%)	0.4086 (+12%)	0.4500 (+11%)	0.4382 (+9.5%)
<i>QE-BM25</i>	0.3097 (+32%)	0.3365 (+31%)	0.4892 (+34%)	0.4882 (+20%)	0.4029 (+1%)

To achieve our evaluation objectives runs demonstrated in Table 2 are conducted. The *QE-DFR* run is conducted to evaluate the performance of our approach compared to a well-known query expansion approach. The *QE-SQC-DFR* run is conducted to evaluate the effectiveness of query expansion via query context modeling using query domain ontology and the impact of semantic relatedness measure in improving retrieval performance. The *RR-SQC-DFR* run is performed to evaluate the effectiveness of re-ranking via query context modeling using query domain ontology and the impact of semantic relatedness measure in improving retrieval performance. The *QE-RR-DFR* run

is conducted to evaluate the effectiveness of query expansion and re-ranking using the query context. The *QE-RR-DFR-Rocc* run is performed to show the effectiveness of our approach when combined with a well-known query expansion method.

Table 2 Performance evaluation of semantic query context based on query domain ontology; results are compared to *BL-TermBased-DFR*

Runs	MAP	Rprec	Bpref	P@5	P@10
<i>BL-TermBased-DFR</i>	0.3017	0.3309	0.426	0.5	0.4853
<i>QE-DFR</i>	0.3332 (+10%)	0.3585 (+8%)	0.4539 (+6.5%)	0.5353 (+7%)	0.5 (+3%)
<i>QE-SQC-DFR</i>	0.3169 (+5%)	0.3508 (+6%)	0.4382 (+2.5%)	0.5588 (+11%)	0.5176 (+6.5)
<i>RR-SQC-DFR</i>	0.3068 (+1.6%)	0.3405 (+3%)	0.433 (+1.6%)	0.5471 (+9.5%)	0.5088 (+4.8%)
<i>QE-RR-DFR</i>	0.3232 (+7%)	0.3623 (+9.5%)	0.437 (+2.5)	0.5706 (+14%)	0.5294 (+9%)
<i>QE-RR-DFR-Rocc</i>	0.3986* (+32%)	0.406* (+22%)	0.5595* (+31%)	0.6882* (+37%)	0.5971* (+23%)

8. Analysis and Discussion

In this chapter we discuss and analyze the influence of conceptual representation of documents and queries, the influence of incorporating keywords dependency in retrieval framework and the impact of semantic query context modeling using association rule mining and query domain ontology on retrieval performance.

8.1. Impact of Conceptual Representation on Retrieval Performance

Keyword-based information retrieval models could return imprecise and erroneous results due to their dependency on specific keywords. Concept-based representation approaches are intended to overcome this limitation by representing queries and documents with semantic concepts using biomedical resources such as UMLS. In this section we focus on evaluating the effectiveness of conceptual representation of documents and queries over keyword-based representation of the collection. Our goal is to investigate the effectiveness of concept-based representation in presenting

information need and how concept-based representation can affect the retrieval performance.

In order to investigate the influence of conceptual representation on retrieval performance the following runs are conducted: (1) *BL-TermBased-BM25* (2) *BL-ConceptBased-BM25* and (3) *BL-Terms-Concepts-BM25*. Figure 16 demonstrates TREC's official performance measures of these runs.

As mentioned before, concept-based index is developed by mapping original documents and queries keywords to concepts in the UMLS using MetaMap. Thus each document and query is represented as a set of UMLS CUIs rather than original keywords.

As Figure 14 shows the performance of the *BL-ConceptBased-BM25* run is decreased for some of the queries compared to the *BL-TermBased-BM25* run. In total the performance of the *BL-ConceptBased-BM25* run has declined by 24% in MAP, 11% in Rprec and 14% in P@5 but the performance has improved in terms of Bpref by 2.5% compared to the *BL-TermBased-BM25* run. We believe that this decline is highly due to the quality of concept extraction and concept mapping of Metamap.

The results of concept-based runs depend on the quality of concept mapping that is provided by MetaMap. Sometimes in the process of mapping, the actual meaning of the original queries or documents is lost and this can degrade the performance. The worst performing queries are examples of this issue. For example query 114 "Adult patients discharged home with palliative care / home hospice" and query 119 "Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent

diabetes” are among the worst concept-based performing queries. MetaMap maps the query concepts “palliative care”, “home hospice” and “anion gap acidosis” to their unique CUIs if only the terms occur adjacent to each other. For example if the term “care” occurs nonadjacent to the term “palliative” MetaMap will map them to two different CUIs. However some of the relevant documents to these queries contain query keywords separately or partly therefore MetaMap is not able to map them to the correct CUIs and consequently the performance decreases.

However as Figure 15 shows, the performance of the *BL-Terms-Concepts-BM25* run is increased compared to *BL-TermBased-BM25* for most of the queries. The *BL-Terms-Concepts-BM25* run improves MAP by 25%, Rprec by 21%, bpref by 19.5%, P@5 by 10% and P@10 by 3% compared to the *BL-TermBased-BM25* run. This demonstrates the positive impact of using both concepts and keywords for query and document representation on the retrieval performance. Using both concepts and terms for representing documents and queries allows us to find more relevant documents in the search results than using only terms. Results demonstrate that concepts have the ability to capture those relevant documents that haven’t been captured with keyword-based representation of documents and queries.

Figure 16 compares the performance of the *BL-TermBased-BM25* run, the *BL-ConceptBased-BM25* run and the *BL-Terms-Concepts-BM25* run. The figure reveals that the *BL-Terms-Concepts-BM25* run improves the performance in all the performance measures compared to both *BL-TermBased-BM25* and *BL-ConceptBased-BM25*.

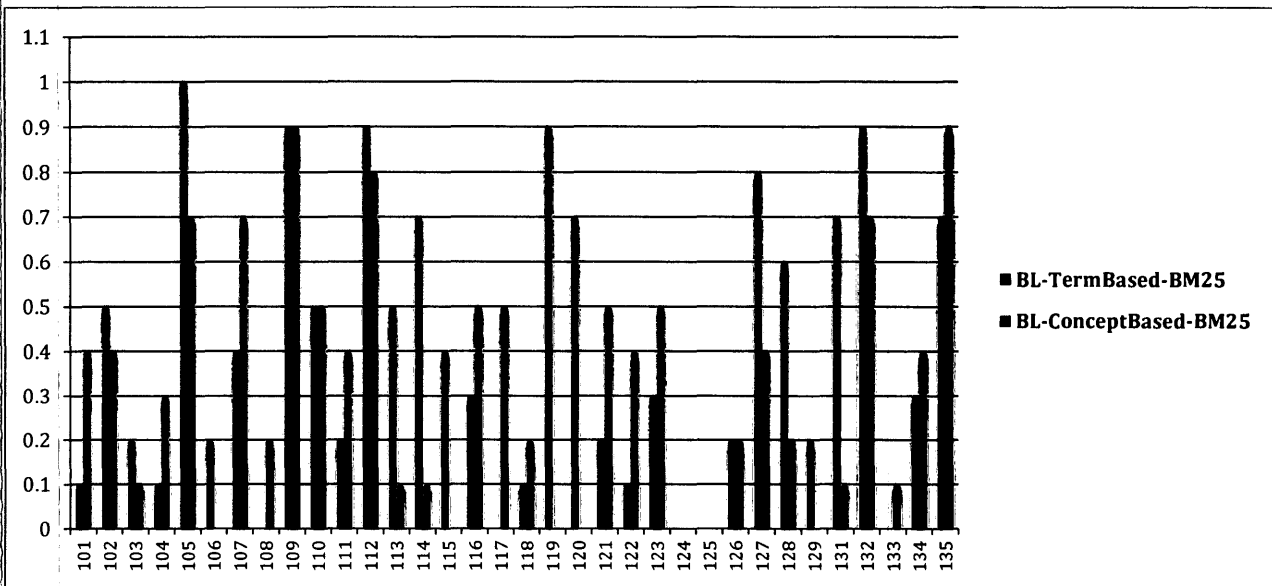


Figure 14. Performance of the BL-TermBased-BM25 run vs performance of the BL-ConceptBased-BM25 run; X-axis indicates the query number and Y-axis indicates the value of the performance measure

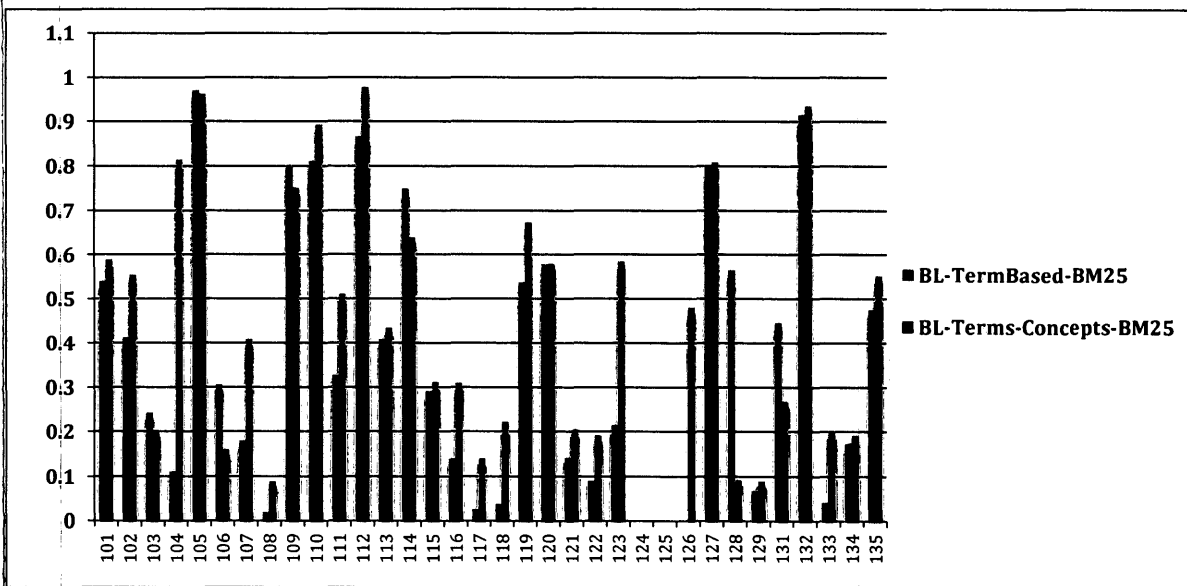


Figure 15. Performance of the BL-TermBased-BM25 run vs performance of the BL-Terms-Concepts_BM25; X-axis indicates the query number and Y-axis indicates the value of the performance measure

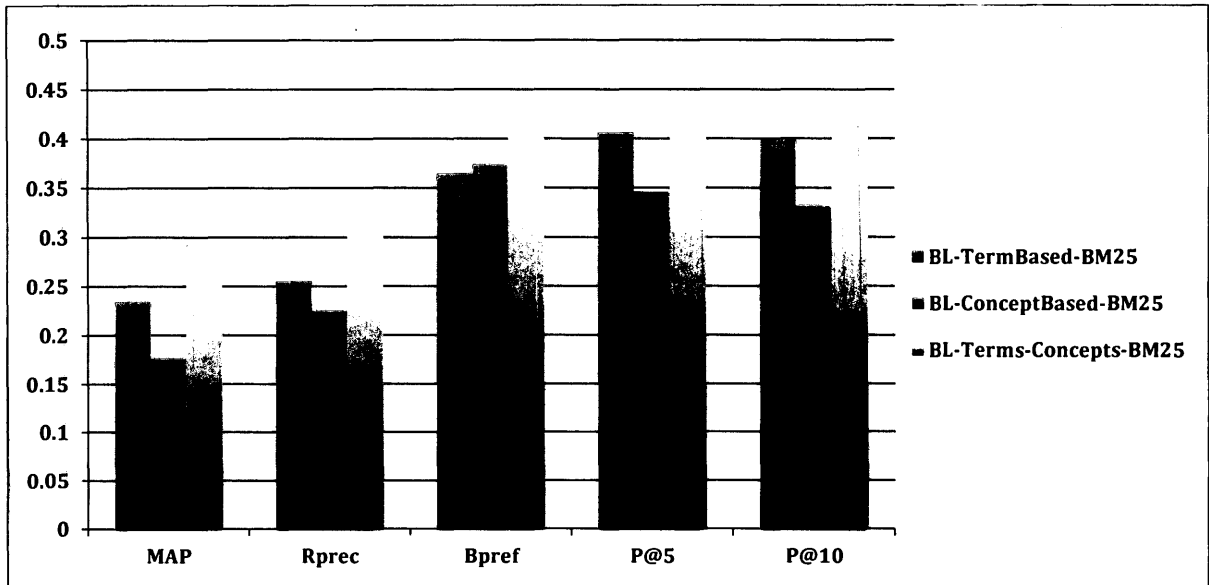


Figure 16. Comparison of performance measures of BL-TermBased-BM25, BL-ConceptBased-BM25 and BL-Terms-Concepts-BM25

8.2. Impact of Incorporating Keywords Dependency in Retrieval Framework

To examine the impact of considering the order of the keywords and their relationships or dependencies we perform the standard term-based retrieval using DFR-based (Divergence from Randomness) dependence model. The DFR-based dependence model incorporate keyword dependency in the retrieval framework. In our experiments we use sequential dependency (SD), which assumes that only neighbor query terms are dependent. In our experimental settings sequential dependency found to be more effective than full dependency. Comparing the results retrieved from DFR-based dependence model (BL-TermBased-DFR) to the best baseline run obtained from BM25 model (BL-Terms-Concepts-BM25) we can confirm that incorporating keywords dependency in to retrieval framework significantly improves the performance. We believe that this is due to

considering the dependency and occurrences of certain pair of terms into the retrieval framework, which is neglected in BM25 probabilistic model.

Figure 17 represents comparison of performance measures of *BL-TermBased-DFR*, *BL-TermBased-BM25*, *BL-ConceptBased-BM25* and *BL-Terms-Concepts-BM25* and reveals that the *BL-TermBased-DFR* run outperforms all the other runs. The *BL-TermBased-DFR* run improves MAP by 3%, Rprec by 7%, P@5 by 11.8% and P@10 by 17.8% compared to *BL-Terms-Concepts-BM25*.

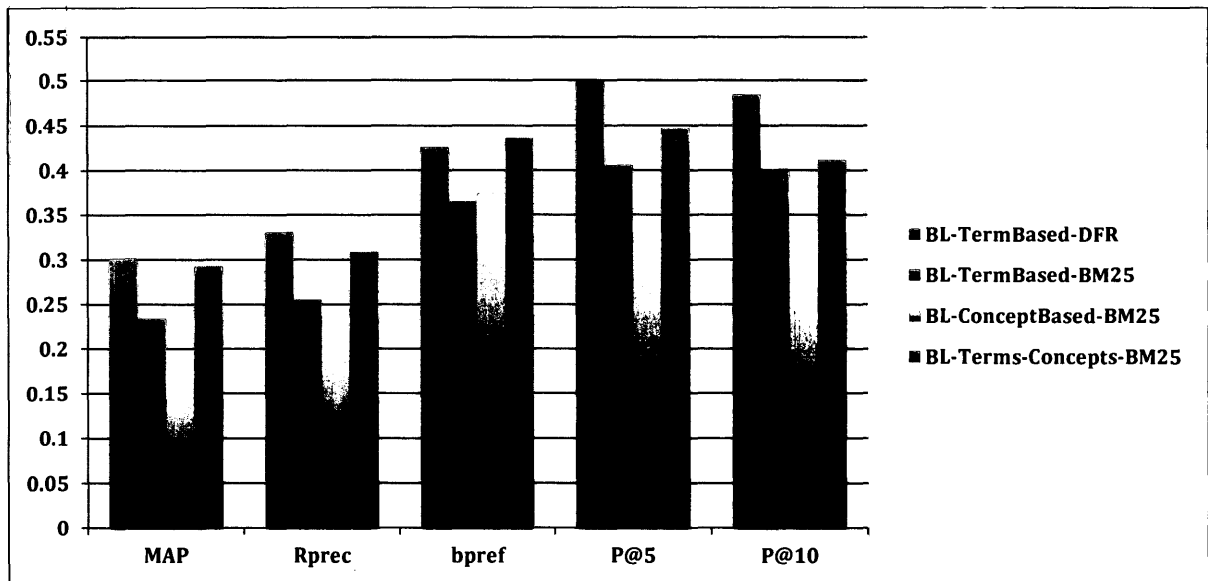


Figure 17. Comparison of performance measures of *BL-TermBased-DFR*, *BL-TermBased-BM25*, *BL-ConceptBased-BM25* and *BL-Terms-Concepts-BM25*

8.3. Impact of Semantic Query Context on Retrieval Performance

In this section we discuss and analyze the impact of query context modeling using association rule mining and query domain ontology. This section also discusses details of the significance test.

8.3.1. Impact of Semantic Association Rule Mining on Retrieval Performance

In order to investigate the influence of using AR mining and semantic relatedness for query context modeling and also to compare our approach to a well-known query expansion method the following runs are performed: *ARM-Supp-Top40Concepts*, *ARM-SQC-Top10Concepts*, *Naïve-Top10Concepts* and *QE-BM25*.

8.3.1.1. Impact of AR Mining on Query Context Modeling

As mentioned before, the *ARM-Supp-Top40Concepts* run relies on re-ranking the initial results according to the query context, where the query context consists of the consequent concepts of those rules that cover the query concepts in their antecedents. These concepts are weighted according to their support. The query context is composed of top 40 retrieved concepts (found as optimal value for this run). As Table 1 shows, this run improves the retrieval performance in all the performance measures. This run improves MAP by 13%, Rprec by 19%, bpref by 18.5%, P@5 by 11.5% and P@10 by 5% compared to *BL-TermBased-BM25*. Results of this run shows that our method of using

AR mining for query context modeling is capable of extracting concepts that are associated with the query concepts. Results of this run confirm the effectiveness of using association rule mining for improving the retrieval performance.

8.3.1.2. Impact of Semantic Relatedness on Query Context Modeling

The naïve method for concept extraction is based on extracting all the concepts from top 20 documents for each query and weight and rank them according to their relatedness to the query concepts. The *Naïve-Top10Concepts* run relies on re-ranking the initial search results using top 10 concepts weighted according to their relatedness to the query concepts. As Table 1 shows, this run improves the retrieval performance in all the performance measures. This run improves MAP by 10%, Rprec by 19%, bpref by 12%, P@5 by 11% and P@10 by 9.5% compared to *BL-TermBased-BM25*. The improvement is due to taking semantic relatedness into consideration even though for query context modeling a naïve method as opposed to AR mining is used to extract concepts. Results can confirm the effectiveness of using semantic relatedness measure in the query context modeling. Results show that the semantic relatedness measure is capable of calculating the relatedness of documents' concepts and query concepts. Results also suggest that semantic relatedness measure is an effective weighting schema to weight concepts in the query context.

8.3.1.3. Impact of Using AR Mining and Semantic Relatedness for Query Context Modeling

The *ARM-SQC-Top10Concepts* run relies on re-ranking the initial search results using query context that is composed of top 10 retrieved concepts from applying TopKRules algorithm to the top 20 documents retrieved from the initial search. Concepts in the query context are weighted according to their relatedness to the query concepts. As Table 1 shows, this run improves the retrieval performance in all the performance measures. This run improves MAP by 19.5%, Rprec by 14%, bpref by 37%, P@5 by 13% and P@10 by 7.5% compared to *BL-TermBased-BM25*. We believe that the improvement is due to the fact that this method has semantic and statistic capabilities meaning that it takes both frequency of concepts and their semantic information into consideration. This run outperforms both *ARM-Supp-Top40Concepts* run and *Naïve-Top10Concepts* run in terms of MAP, bpref and P@5. It also outperforms *ARM-Supp-Top40Concepts* and is comparable to *Naïve-Top10Concepts* in terms of P@10. However the performance has degraded by 4% in terms of Rprec compared to *ARM-Supp-Top40Concepts* and *Naïve-Top10Concepts*. Results of this run confirm the effectiveness of semantic query context modeling using association rule mining for concept extraction and semantic relatedness measure for concept weighting.

The *QE-BM25* run is based on applying Rochhio's query expansion algorithm on *BL-Terms-Concepts-BM25*. The results of the *QE-BM25* run shows that this run is effective for improving the retrieval performance by 32% at MAP, 31% at Rprec, 34% at bpref, 20% at P@5 and 1% at P@10.

Figure 18 compares the performance measures of *ARM-Supp-Top40Concepts*, *ARM-SQC-Top10Concepts*, *Naive-Top10Concepts* and *QE-BM25*. Although the figure shows that the *QE-BM25* run performs better than our proposed runs in terms of MAP, Rprec and P@5 but our proposed approach (*ARM-Supp-Top40Concepts*) performs better in terms of bpref and P@10. Note that the *QE-BM25* is applied on concepts and terms index, which performs better than term-based index. However our proposed approach is based on re-ranking initial results from term-based index. Overall, our proposed approach is promising. This is because the results generated from our approach significantly (statistical significance for our approach is denoted * over *BL-TermBased-BM25* in Table 1 details are explained in section 8.3.6) improve the performance compared to the baseline run at all the TREC's official performance measures especially at bpref.

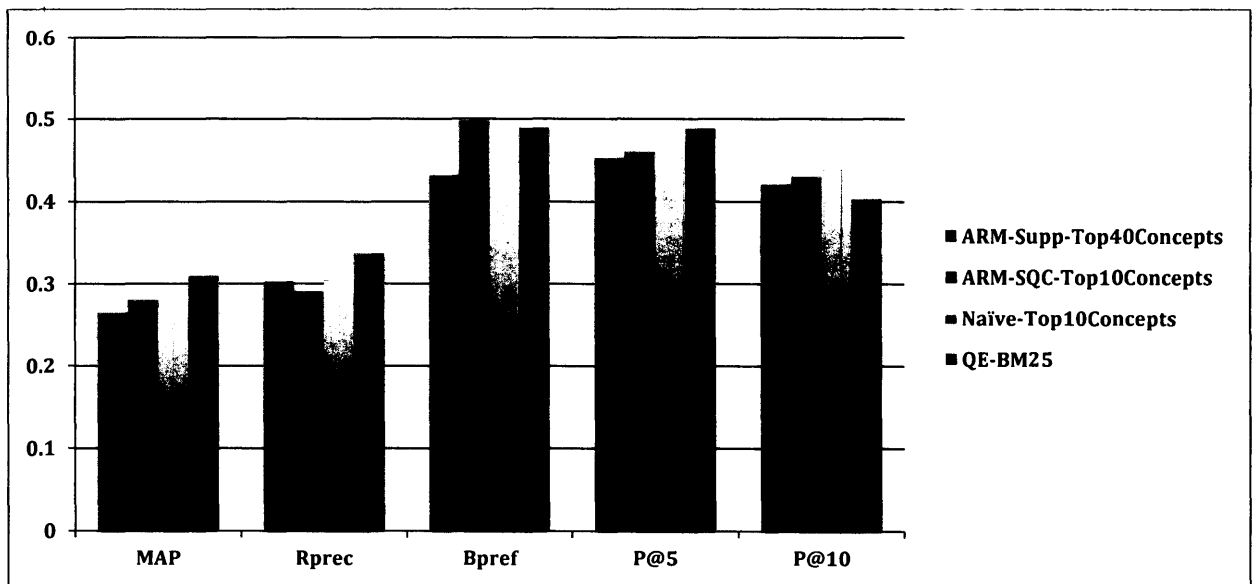


Figure 18. Comparison of performance measures of *ARM-Supp-Top40Concepts*, *ARM-SQC-Top10Concepts*, *Naive-Top10Concepts* and *QE-BM25*

8.3.2. Impact of Query Domain Ontology on Retrieval Performance

In order to investigate the influence of using query domain ontology for modeling a semantic query context and also to compare our approach to a well-known query expansion method the following runs are conducted: *QE-SQC-DFR*, *RR-SQC-DFR*, *QE-RR-DFR*, *QE-RR-DFR-Rocc* and *QE-DFR*. In this section we investigate the influence of re-ranking of initial results and query expansion using semantic query context. The semantic query context consists of concepts in the query domain ontology that are weighted according to their relatedness to the original query concepts.

8.3.2.1. Impact of Query Expansion Using Semantic Query Context

As mentioned before, the *QE-SQC-DFR* run relies on expanding the original queries using the semantic query context, where we use weight threshold of 0.9. As Table 2 shows, this run improves the retrieval performance in all the performance measures compared to *BL-TermBased-DFR*. This run improves MAP by 5%, Rprec by 6%, bpref by 2.5%, P@5 by 11% and P@10 by 6.5% compared to *BL-TermBased-DFR*. Results of this run suggest the effectiveness expanding the original queries using query domain ontology and ability of medical query domain ontology for extracting associated diseases, treatments and symptoms. Expanding the original queries using highly related and associated concepts and weight them according to their relatedness to the original concepts boosts the retrieval performance by capturing those relevant documents not

containing original queries' keywords explicitly. The positive results also indicate the usefulness of combining semantic search with standard retrieval model.

8.3.2.2. Impact of Re-ranking Using Semantic Query Context

The *RR-SQC-DFR* run relies on re-ranking the initial search results based on the semantic query context, where the initial search results are retrieved from the *BL-TermBased-DFR* run. As Table 2 shows, this run improves the retrieval performance in all the performance measures compared to the *BL-TermBased-DFR* run. This run improves MAP by 1.6%, Rprec by 3%, bpref by 1.6%, P@5 by 9.5% and P@10 by 4.8% compared to *BL-TermBased-DFR*. Results of this run suggest the effectiveness of re-ranking using semantic query context. Re-ranking the initial search results using the semantic query context boosts the retrieval performance by giving a new score to the retrieved documents. The new score is assigned by considering those concepts that are highly related to the original query concepts. Although this run improves the retrieval performance in all the performance measure compared to the *BL-TermBased-DFR* run, but the *QE-SQC-DFR* run outperforms this run in all the performance measures. We believe that this is due to the ability of the *QE-SQC-DFR* run in capturing new relevant documents as opposed to the *RR-SQC-DFR* run, where new documents are not extracted but new scores are assigned to the initial search results.

8.3.3. Impact of Combining Query Expansion and Re-ranking Using Semantic Query Context

Since the *QE-SQC-DFR* run outperforms the *RR-SQC-DFR* run in all the performance measures we conduct the *QE-RR-DFR* run. This run relies on re-ranking the results obtained from *QE-SQC-DFR* run. Basically for this run first we expand the original queries using the semantic query context then we re-rank the results retrieved from semantic query expansion according to the semantic query context. As Table 2 shows, this run improves the retrieval performance in all the performance measures compared to *BL-TermBased-DFR*. This run improves MAP by 7%, Rprec by 9.5%, bpref by 2.5%, P@5 by 14% and P@10 by 9% compared to *BL-TermBased-DFR*. Moreover, this run outperforms both *QE-SQC-DFR* and *RR-SQC-DFR*. This run improves MAP by 1.9%, Rprec by 3.2%, P@5 by 2.1% and P@10 by 2.2% compared to *QE-SQC-DFR*. This run also improves MAP by 5.3%, Rprec by 6.4, P@5 by 4.2% and P@10 by 4% compared to *RR-SQC-DFR*. Figure 19 shows comparison of performance measures of *QE-SQC-DFR*, *RR-SQC-DFR* and *QE-RR-DFR* and reveals that the *QE-RR-DFR* run outperforms the other two runs.

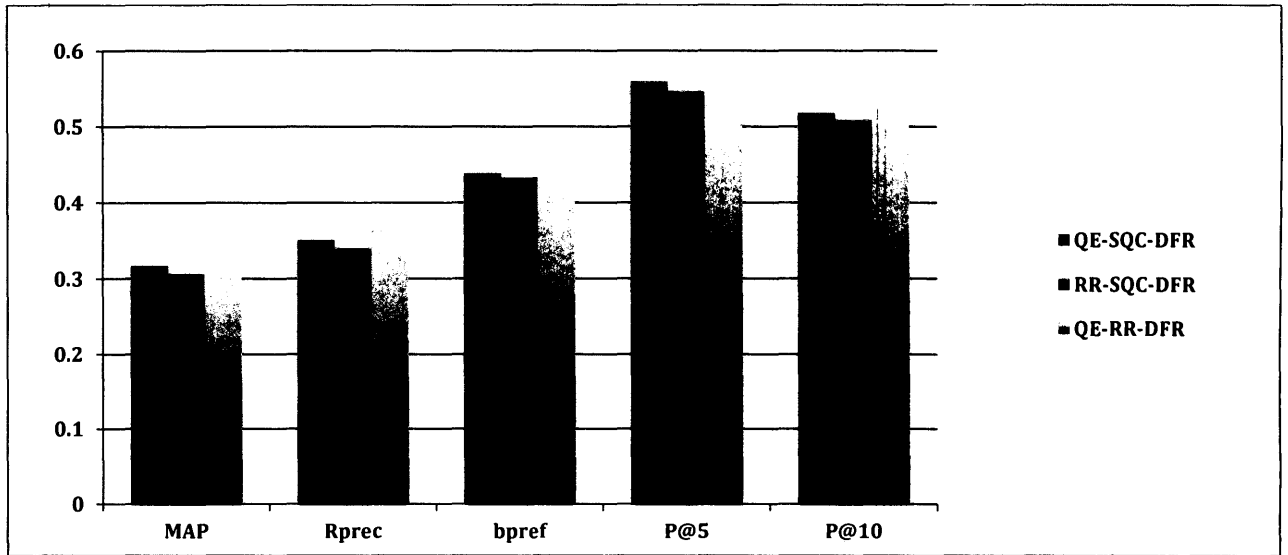


Figure 19. Comparison of performance measures of QE-SQC-DFR, RR-SQC-DFR and QE-RR-DFR

8.3.4. Impact of Rocchio's Query Expansion Mechanism

The *QE-DFR* run is performed using Rocchio's relevance feedback algorithm where we use Terrier's default value of 10 as the number of terms to expand a query and the default value of 3 as the number of top-ranked documents. As Table 2 shows, this run improves the retrieval performance in all the performance measures compared to *BL-TermBased-DFR*. This run improves MAP by 10%, Rprec by 8%, bpref by 6.5%, P@5 by 7% and P@10 by 3% compared to *BL-TermBased-DFR*. However our proposed approach (*QE-RR-DFR*) outperforms this run in Rprec (by 1%), P@5 (by 6.5%) and P@10 (by 5.8%) and is comparable to Rocchio's query expansion mechanism in terms of bpref and MAP.

8.3.5. Impact of integrating Rocchio into Semantic Query Context

To evaluate the effectiveness of our proposed approach after applying Rocchio's query expansion mechanism, we conduct the *QE-RR-DFR-Rocc* run. This run relies on re-ranking the results obtained from applying Rocchio's algorithm on *QE-SQC-DFR* run. Basically, for this run first we expand the original queries using the semantic query context and Rocchio's query expansion mechanism, then we re-rank the results retrieved from query expansion according to the semantic query context. As Table 2 shows, this run improves the retrieval performance in all the performance measures compared to *BL-TermBased-DFR*. This run improves MAP by 32%, Rprec by 22%, bpref by 31%, P@5 by 37% and P@10 by 23% compared to *BL-TermBased-DFR*. This run improves MAP by 23%, Rprec by 12%, bpref by 28%, P@5 by 20% and P@10 by 12% compared to *QE-RR-DFR*. This run also improves MAP by 19%, Rprec by 13%, bpref by 23%, P@5 by 28% and P@10 by 19% compared to *QE-DFR*. Figure 20 represents comparison of performance measures of *QE-RR-DFR*, *QE-RR-DFR-Rocc* and *QE-DFR* and reveals that the *QE-RR-DFR-Rocc* run outperforms the other two runs. Results from this run foster the effectiveness of integrating semantic query context into Rocchio's query expansion mechanism. The effectiveness of this approach is due to considering the frequency of keywords in top retrieved documents and the semantic information in the retrieval framework.

Overall, our proposed approach is promising. This is because the results generated from our approach significantly (statistical significance for our approach is denoted * over *BL-*

TermBased-DFR in Table 2, details are explained in section 8.3.6) improve the performance compared to the baseline run and the Rocchio query expansion run at all the TREC's official performance measures especially at P@5. Figure 20 represents comparison of performance measure of *QE-RR-DFR*, *QE-RR-DFR-Rocc* and *QE-DFR* and reveals that the *QE-RR-DFR-Rocc* run outperforms the other two runs in all the performance measures.

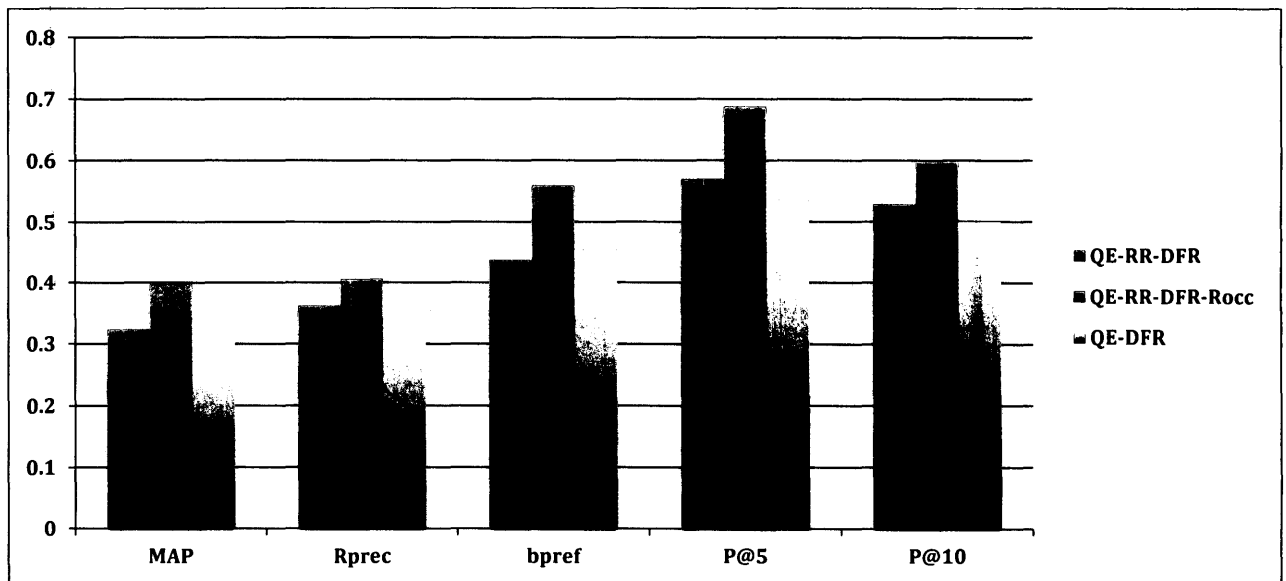


Figure 20. Comparison of performance measures of *QE-RR-DFR*, *QE-RR-DFR-Rocc* and *QE-DFR*

8.3.6. Significance test

Our main goal in the IR field is to find better retrieval approaches and prevent the promotion of methods that perform better by chance. Although TREC offers standard evaluation tool but we know that sometimes there is undeniable noise in these evaluations. That's due to the fact that some topics are harder than others. Moreover the

assessors that judge the degree of relevance of the documents are open to variability in their behavior. Statistical significance tests can aid us in detecting significance improvements of one approach over the other [52]. In order to prove that our proposed approaches truly perform better than the baseline and they are not performing better only by chance, we perform significant tests. We use the Wilcoxon Signed-Rank test to evaluate the difference between our proposed approach and the baseline. Table 3, Table 4, Table 5, Table 6, Table 7, Table 8, Table 9, Table 10, Table 11 and Table 12 report the results of our significance tests. As the tables show, modeling query context via association rule mining improves the performance significantly at MAP and bpref measures and modeling query context using query domain ontology improves the performance significantly at all the TREC's official measures.

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -1.7866. The p-value is 0.03673. The result is not significant at $p \leq 0.01$.	The Z-value is -1.7866. The p-value is 0.07346. The result is not significant at $p \leq 0.01$.
	The W-value is 193. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 193. The distribution is approximately normal. Therefore, the Z-value above should be used.
Significance level =0.05	The Z-value is -1.7866. The p-value is 0.03673. The result is significant at $p \leq 0.05$.	The Z-value is -1.7866. The p-value is 0.07346. The result is not significant at $p \leq 0.05$.
	The W-value is 193. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 193. The distribution is approximately normal. Therefore, the Z-value above should be used.

Table 3. Results of significance test for *ARM-SQC-Top10Concepts* run (MAP measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -2.3962. The p-value is 0.0082. The result is significant at $p \leq 0.01$.	The Z-value is -2.3962. The p-value is 0.0164. The result is not significant at $p \leq 0.01$.
	The W-value is 116. The critical value of W for N = 30 at $p \leq 0.01$ is 120. Therefore, the result is significant at $p \leq 0.01$.	The W-value is 116. The critical value of W for N = 30 at $p \leq 0.01$ is 109. Therefore, the result is not significant at $p \leq 0.01$.
Significance level =0.05	The Z-value is -2.3962. The p-value is 0.0082. The result is significant at $p \leq 0.05$.	The Z-value is -2.3962. The p-value is 0.0164. The result is significant at $p \leq 0.05$.
	The W-value is 116. The critical value of W for N = 30 at $p \leq 0.05$ is 151. Therefore, the result is significant at $p \leq 0.05$.	The W-value is 116. The critical value of W for N = 30 at $p \leq 0.05$ is 137. Therefore, the result is significant at $p \leq 0.05$.

Table 4. Results of significance test for *ARM-SQC-Top10Concepts* run (bpref measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -1.068. The p-value is 0.14231. The result is not significant at $p \leq 0.01$.	The Z-value is -1.068. The p-value is 0.28462. The result is not significant at $p \leq 0.01$.
	The W-value is 193.5. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 193.5. The distribution is approximately normal. Therefore, the Z-value above should be used.
Significance level =0.05	The Z-value is -1.068. The p-value is 0.14231. The result is not significant at $p \leq 0.05$.	The Z-value is -1.068. The p-value is 0.28462. The result is not significant at $p \leq 0.05$.
	The W-value is 193.5. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 193.5. The distribution is approximately normal. Therefore, the Z-value above should be used.

Table 5. . Results of significance test for run *ARM-SQC-Top10Concepts* run (R-prec measur

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -0.504. The p-value is 0.30854. The result is not significant at $p \leq 0.01$.	he Z-value is -0.504. The p-value is 0.61708. The result is not significant at $p \leq 0.01$.
	The W-value is 91.5. The critical value of W for N = 20 at $p \leq 0.01$ is 43. Therefore, the result is not significant at $p \leq 0.01$.	The W-value is 91.5. The critical value of W for N = 20 at $p \leq 0.01$ is 37. Therefore, the result is not significant at $p \leq 0.01$.
Significance level =0.05	The Z-value is -0.504. The p-value is 0.30854. The result is not significant at $p \leq 0.05$.	The Z-value is -0.504. The p-value is 0.61708. The result is not significant at $p \leq 0.05$.
	The W-value is 91.5. The critical value of W for N = 20 at $p \leq 0.05$ is 60. Therefore, the result is not significant at $p \leq 0.05$.	The W-value is 91.5. The critical value of W for N = 20 at $p \leq 0.05$ is 52. Therefore, the result is not significant at $p \leq 0.05$.

Table 6. Results of significance test for *ARM-SQC-Top10Concepts* run (P@5 measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -1.2243. The p-value is 0.11123. The result is not significant at $p \leq 0.01$.	The Z-value is -1.2243. The p-value is 0.22246. The result is not significant at $p \leq 0.01$.
	The W-value is 117. The critical value of W for N = 25 at $p \leq 0.01$ is 76. Therefore, the result is not significant at $p \leq 0.01$.	The W-value is 117. The critical value of W for N = 25 at $p \leq 0.01$ is 68. Therefore, the result is not significant at $p \leq 0.01$.
Significance level =0.05	The Z-value is -1.2243. The p-value is 0.11123. The result is not significant at $p \leq 0.05$.	The Z-value is -1.2243. The p-value is 0.22246. The result is not significant at $p \leq 0.05$.
	The W-value is 117. The critical value of W for N = 25 at $p \leq 0.05$ is 100. Therefore, the result is not significant at $p \leq 0.05$.	The W-value is 117. The critical value of W for N = 25 at $p \leq 0.05$ is 89. Therefore, the result is not significant at $p \leq 0.05$.

Table 7. Results of significance test for *ARM-SQC-Top10Concepts* run (P@10 measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -3.1358. The p-value is 0.00084. The result is significant at $p \leq 0.01$.	The Z-value is -3.1358. The p-value is 0.00168. The result is significant at $p \leq 0.01$.
	The W-value is 105. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 105. The distribution is approximately normal. Therefore, the Z-value above should be used.
Significance level =0.05	The Z-value is -3.1358. The p-value is 0.00084. The result is significant at $p \leq 0.05$.	The Z-value is -3.1358. The p-value is 0.00168. The result is significant at $p \leq 0.05$.
	The W-value is 105. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 105. The distribution is approximately normal. Therefore, the Z-value above should be used.

Table 8. Results of significance test for *QE-RR-DFR-Rocc* run (MAP measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -4.3853. The p-value is 0. The result is significant at $p \leq 0.01$.	The Z-value is -4.3853. The p-value is 0. The result is significant at $p \leq 0.01$.
	The W-value is 41. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 41. The distribution is approximately normal. Therefore, the Z-value above should be used.
Significance level =0.05	The Z-value is -4.3853. The p-value is 0. The result is significant at $p \leq 0.05$.	The Z-value is -4.3853. The p-value is 0. The result is significant at $p \leq 0.05$.
	The W-value is 41. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 41. The distribution is approximately normal. Therefore, the Z-value above should be used.

Table 9. Results of significance test for *QE-RR-DFR-Rocc* run (bpref measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -2.6072. The p-value is 0.00453. The result is significant at $p \leq 0.01$.	The Z-value is -2.6072. The p-value is 0.00906. The result is significant at $p \leq 0.01$.
	The W-value is 145. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 145. The distribution is approximately normal. Therefore, the Z-value above should be used.
Significance level =0.05	The Z-value is -2.6072. The p-value is 0.00453. The result is significant at $p \leq 0.05$.	The Z-value is -2.6072. The p-value is 0.00906. The result is significant at $p \leq 0.05$.
	The W-value is 145. The distribution is approximately normal. Therefore, the Z-value above should be used.	The W-value is 145. The distribution is approximately normal. Therefore, the Z-value above should be used.

Table 10. Results of significance test for *QE-RR-DFR-Rocc* run (R-prec measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -3.3246. The p-value is 0.00045. The result is significant at $p \leq 0.01$.	The Z-value is -3.3246. The p-value is 0.0009. The result is significant at $p \leq 0.01$.
	The W-value is 57. The critical value of W for N = 28 at $p \leq 0.01$ is 101. Therefore, the result is significant at $p \leq 0.01$.	The W-value is 57. The critical value of W for N = 28 at $p \leq 0.01$ is 91. Therefore, the result is significant at $p \leq 0.01$.
Significance level =0.05	The Z-value is -3.3246. The p-value is 0.00045. The result is significant at $p \leq 0.05$.	The Z-value is -3.3246. The p-value is 0.0009. The result is significant at $p \leq 0.05$.
	The W-value is 57. The critical value of W for N = 28 at $p \leq 0.05$ is 130. Therefore, the result is significant at $p \leq 0.05$.	The W-value is 57. The critical value of W for N = 28 at $p \leq 0.05$ is 116. Therefore, the result is significant at $p \leq 0.05$.

Table 11. Results of significance test for *QE-RR-DFR-Rocc* run (P@5 measure)

	One-tailed	Two-tailed
Significance level= 0.01	The Z-value is -2.8806. The p-value is 0.00199. The result is significant at $p \leq 0.01$.	The Z-value is -2.8806. The p-value is 0.00398. The result is significant at $p \leq 0.01$.
	The W-value is 76.5. The critical value of W for N = 28 at $p \leq 0.01$ is 101. Therefore, the result is significant at $p \leq 0.01$.	The W-value is 76.5. The critical value of W for N = 28 at $p \leq 0.01$ is 91. Therefore, the result is significant at $p \leq 0.01$.
Significance level =0.05	The Z-value is -2.8806. The p-value is 0.00199. The result is significant at $p \leq 0.05$.	The Z-value is -2.8806. The p-value is 0.00398. The result is significant at $p \leq 0.05$.
	The W-value is 76.5. The critical value of W for N = 28 at $p \leq 0.05$ is 130. Therefore, the result is significant at $p \leq 0.05$.	The W-value is 76.5. The critical value of W for N = 28 at $p \leq 0.05$ is 116. Therefore, the result is significant at $p \leq 0.05$.

Table 12. Results of significance test for *QE-RR-DFR-Rocc* run (P@10 measure)

8.4. Comparison of Approaches

In this section we focus on comparing our proposed approaches. The first proposed approach involves modeling a query context using association rule mining where concepts in the query context are weighted according to their semantic relatedness to the original query concepts. The *ARM-SQC-Top10Concepts* run is representative of our first approach. This run significantly improves the performance compared to the baseline run. On the other hand our second proposed approach involves modeling a query context using query domain ontology where concepts in the query context are weighted according to their semantic relatedness to the original query concepts. The *QE-RR-DFR* run is representative of our second approach. This run also significantly improves the performance compared to the baseline run. However our second approach of query

context modeling (*QE-RR-DFR*) outperforms our first approach (*ARM-SQC-Top10Concepts*) at all the TREC's official measures. The *QE-RR-DFR* run improves MAP by 15%, Rprec by 19%, bpref by 6.9%, P@5 by 26% and P@10 by 20% compared to the *ARM-SQC-Top10Concepts*. Results show that modeling a query context using query domain ontology is more effective than modeling a query context using AR mining. Results indicate that the query domain ontology is more capable in expressing closely related and highly associated concepts compared to AR mining. This is due to the fact that AR mining algorithms extract concepts that are statistically significant to the source dataset. Although we weight the extracted concepts semantically but they are initially extracted based on their frequency in the source dataset. However the query domain ontology consists of those concepts that have a semantic relation with the original query concepts. Therefore concepts in the query context are extracted semantically (not statistically) and are also weighted semantically. As a result a query context that is modeled based on query domain ontology consists of more semantically related concepts to the original query than a query context that is modeled based on AR mining.

Figure 21 compares performance measures of the *ARM-SQC-Top10Concepts* run and the *QE-RR-DFR* run and reveals that the *QE-RR-DFR* run outperforms the *ARM-SQC-Top10Concepts* run at all the performance measures. The results generated from the *QE-RR-DFR* run significantly (details are explained in section 8.3.6) improve the performance compared to the *ARM-SQC-Top10Concepts* run.

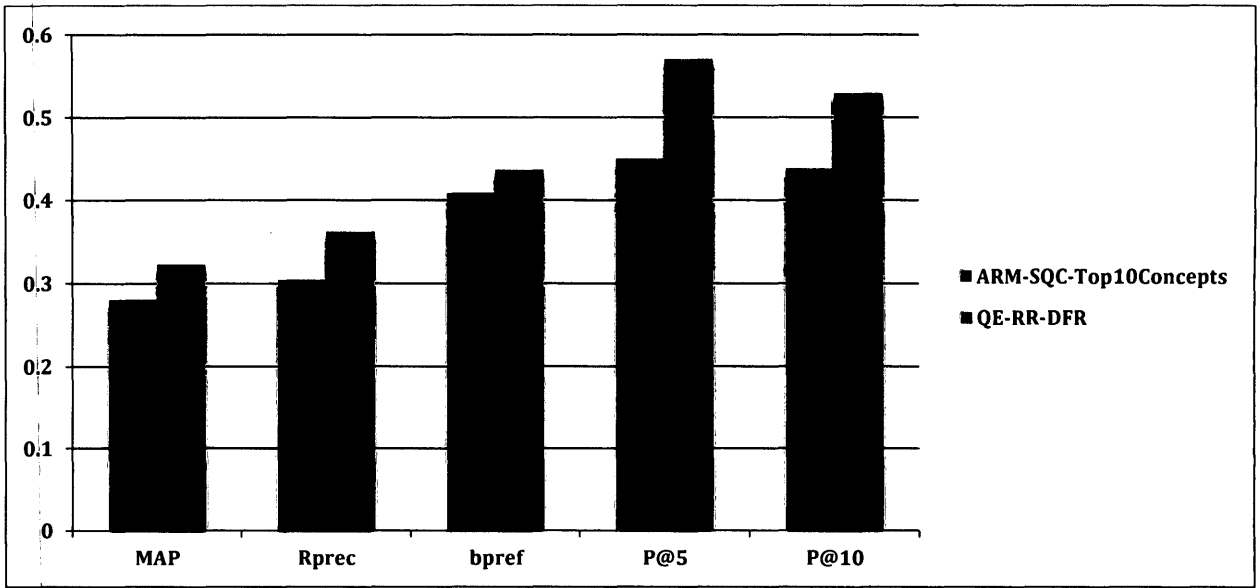


Figure 21. Comparison of performance measures of ARM-SQC-Top10Concepts and QE-RR-DFR

9. Conclusions and Future Work

In this chapter we summarize our contributions on exploiting semantic information for improving clinical text retrieval. First we summarize our contribution on using semantic association rule mining for improving clinical text retrieval and then we summarize our contribution on using semantic query domain ontology for improving clinical text retrieval. Moreover, this chapter discusses conclusions and future work of our work.

9.1. Semantic Association Rule Mining for Improving Clinical Text Retrieval

AR mining has been widely used on the electronic medical records for discovering hidden knowledge and medical patterns. AR mining is also used for improving the information retrieval performance via query reformulation and query expansion. One of the obstacles in AR mining is that often a huge number of rules are generated even with very reasonable support and confidence. The other obstacle in AR mining is the low quality of the generated rules, since many of the generated rules are redundant or semantically wrong or conflict with common sense. Current data mining techniques can efficiently generate association rules that are statistically significant to the source dataset where rules

satisfy the user-specified interestingness measure. Interestingness measures select and rank discovered rules according to user's interest. Current interestingness measures consider frequency of items and time and space cost but they ignore semantics. Therefore the main challenge of using AR in IR is to select the rules that are related to the query using appropriate interestingness measure.

In this study, we propose a novel approach to modeling medical query context based on mining semantic-based AR mining for improving clinical text retrieval. We semantically index the EMRs with concepts of UMLS ontology. First, the concepts in the query context are derived from the rules that cover the query and then weighted according to their semantic relatedness to the query concepts. The query context is then exploited to re-rank patients records for improving clinical retrieval performance.

9.2. Semantic Query Domain Ontology for Improving Clinical Text Retrieval

Clinical IR presents several challenges including terminology mismatch and granularity mismatch. Terminology mismatch includes synonym problem and polysemy problem, where synonymy problem refers to same terms with different meanings and polysemy problem refers to different terms with the same meaning. Terms in queries are sometimes too general however those terms in the relevant documents are too specific or vice versa, this leads to granularity mismatch problem. Moreover, one of the main objectives in clinical IR is to fill the semantic gap among the queries and documents and go beyond keywords matching.

To address these issues, we propose a novel approach to modeling medical query context

based on query domain ontology where we use semantic information to improve the performance of clinical IR systems by representing queries in an expressive and meaningful context. First, we semantically index the EMRs with concepts of UMLS ontology. To model a query context, initially we model and develop query domain ontology. The query domain ontology represents concepts closely related with the query concepts. Query context represents concepts extracted from query domain ontology and weighted according to their semantic relatedness to the query concepts. The query context is then exploited in query expansion and patients' records re-ranking for improving clinical retrieval performance.

9.3. Conclusions

In this study we present two novel approaches for modeling semantic query context. First we present a novel medical query context modeling based on association rule mining and semantic relatedness measures, which is then exploited for re-ranking clinical search results. We evaluate our proposed approach on the challenging ad-hoc retrieval task of TREC Medical Records Track. Results show that our proposed method improves the retrieval performance significantly compared to the baseline (probabilistic BM25 model) search. Our proposed approach also allows performing better retrieval performance compared to naïve methods used for query context modeling. Moreover our proposed approach performs better than the well-known Rocchio's query expansion mechanism in terms of bpref and P@10 and is comparable to the Rocchio's run in terms of other performance measures. Note that the Rocchio's query expansion algorithm is applied on

concepts and terms index, which performs better than term-based index. However our proposed approach is based on re-ranking initial results from term-based index. The results suggest the effectiveness of using semantic relatedness measure as the interestingness measure for selecting and ranking generated rules from applying the AR mining algorithm. Overall, our proposed approach is promising. This is because the results generated from our approach significantly improve the performance compared to the baseline run at all the TREC's official performance measures especially at bpref.

In our second approach we present a novel medical query context modeling based on query domain ontology and semantic relatedness measures, which is then exploited for query expansion and re-ranking clinical search results. We evaluate our proposed approach of modeling semantic query context using semantic query domain ontology on the challenging ad-hoc retrieval task of TREC medical records track. Results show that our proposed approach significantly improves the retrieval performance at all the TREC's official measures compared to the baseline (the DFR-based dependence model) search. Our proposed approach also performs better than the well-known Rocchio's query expansion algorithm at all the TREC's official performance measures especially at P@5. The results suggest the effectiveness of combining conceptual search using query context modeling with standard term-based retrieval. Results also indicate capability of query domain ontology in expressing closely related concepts and effectiveness of using semantic relatedness measure for weighting query context concepts. Overall, our proposed approach is promising. This is because the results generated from our approach significantly improve the performance compared to the baseline run and outperforms the

Rocchio query expansion run at all the TREC's official performance measures especially at P@5.

Comparing our two proposed approaches we can conclude that modeling a query context using query domain ontology is more effective than modeling a query context using AR mining. This is due to the fact that the query domain ontology run outperforms the AR mining run. Therefore we believe that the query domain ontology is more capable in expressing closely related and highly associated concepts compared to AR mining. This is due to the fact that although we weight the extracted concepts from AR mining semantically but they are initially extracted based on their frequency in the source dataset. However the query domain ontology consists of those concepts that have a semantic relation with the original query concepts. As a result concepts in the query context that is based on query domain ontology are extracted semantically and are also weighted semantically as opposed to concepts in the query context that is based on AR mining where concepts are extracted statistically and weighted semantically. Consequently query context that is modeled based on query domain ontology consists of more semantically related concepts to the original query concepts.

9.4. Future Work

The future work will focus on modeling query context using sequential pattern mining and semantic relatedness measures. The objective is to evaluate the impact of identifying temporal patterns using sequential pattern mining and time series mining on the retrieval

performance. We will also focus on finding rule mining techniques that provide an alternative to frequent itemset mining. As opposed to current association rule mining techniques that are based on frequent itemset mining we are planning to propose techniques that are based on user-specified interestingness measures and semantic-based measures.

Furthermore for future work, we plan to do the retrieval process based on calculating the relatedness of query concepts and documents' keywords using definition information and co- occurrence information. In other words we plan to score, retrieve and rank documents based on their semantic relatedness to the queries. Where the score of a document is calculated based on the sum of the semantic relatedness scores of its concepts. We will also examine the impact of exploiting semantics on Medical TREC Records Track 2012 and 2013.

Works Cited

- [1]. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *International Conference on Management of Data* (pp. 207-216). Washington: ACM.
- [2]. Agrwal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. *Intenational Conference on Very Large Databases* (pp. 487-499). Santiago de Chile: Morgan Kaufmann.
- [3]. Amati, G. (2003). *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow.
- [4]. Antonie, M., & R.Zaiane, O. (2004). Mining Positive and Negative Association Rules: An Approach for Confined Rules. *8th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 27-38). Pisa: Springer.
- [5]. Aronson, A., & Lang, F. (2010). An Overview of MetaMap: Historical and Recent Advances. *Journal of the American Medical Informatics Association* , 17 (3), 229-236.
- [6]. Babashzadeh, A., Daoud, M., & Huang, J. (2013). Using Semantic-Based Association Rule Mining for Improving Clinical Text Retrieval. *The 2nd International Conference on Health Information Science* (pp. 186-197). London: Springer.
- [7]. Babashzadeh, A., Huang, J., & Daoud, M. (2013). Exploiting Semantics for Improving Clinical Information Retrieval. *36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin: ACM.

- [8]. Babashzadeh, A., Raissi, H., Daoud, M., & Huang, J. (2012). Query Expansion Using Multiple External Resources and Domain Knowledge for Medical Information Retrieval. *2nd Advances in Health Informatics Conference*. Toronto.
- [9]. Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic Query Expansion Using SMART: TREC 3. *TREC*.
- [10]. Carpineto, C., & Romano, G. (2012). A Survey of Automatic Query Rxpansion in Information Retrieval. *ACM Computing Surveys* , 44 (1), 1-50.
- [11]. D.Manning, C., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.
- [12]. Daoud, M., Kasperowicz, D., Miao, J., & Huang, J. (2011). York University at TREC 2011, Medical Records Track. *TREC*. Maryland.
- [13]. Eichmann, D. (2011). Concept-centric Indexing and Retrieval on Medical Text. *TREC*. Maryland.
- [14]. Fournier-viger, P., Wu, C., & Tseng, V. (2012). Mining Top-K Association Rules. *25th Canadian Conference on Artificial Intelligence* (pp. 61-73). Toronto: Springer.
- [15]. Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The Vocabulary Problem in Human-System Communication. *Communication of the ACM* , 30 (11), 964-971.
- [16]. Gang, L., Zheng, C., & Zhang, L. (2009). Text Information Retrieval based on Concept Semantic Similarity. *5th International Conference on Semantics, Knowledge and Grid*, (pp. 356-360). Zhuhai .
- [17]. Ganguly, P., A.Rabhi, F., & K.Ray, P. (2003). Bridging Semantic Gap. *Proceedings of the 2002 Conference on Pattern Languages of Programs* (pp. 59-61). Darlinghurst: Australian Computer Society.
- [18]. Gobeil, G., Gaudinat, A., Pasche, E., Teodoro, D., Vishnyakova, D., & Ruch , P. (2011). BiTeM Group Report for TREC Medical Records Track 2011. *TREC*.
- [19]. Herch, W., Hickam, D., Haynes, R., & McKibbon, K. (1994). A Performance and Failure Analysis of SAPPHERE with a MEDLINE Test Collection. *Journal of the American Medical Informatics Association* , 1 (1), 51-60.

- [20]. Hirst, G., & St-Onge, D. (1998). *Lexical Chains as Representation of Context for the Detection and Correction of Malapropisms*. (C. Fellbaum, Ed.) Cambridge: MIT press.
- [21]. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E., & Milios, E. (2006). Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems* , 3 (3), 55-73.
- [22]. Huang, X., & Hu, Q. (2009). A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval. *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 307-314). Boston: ACM.
- [23]. Huang, X., An, A., & Hu, Q. (2010). Medical Search and Classification Tools for Recommendation. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 707). Geneva: ACM.
- [24]. Huang, X., Wen, M., An, A., & Hunag, Y. (2006). A Platform for Okapi-based Contextual Information Retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 728). Seattle: ACM.
- [25]. Icev, A., Ruiz, C., & Ryder, E. (2003). Distance-enhanced Association Rules for Gene Expression. *Proceedings of the 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, (pp. 34-40). Washington.
- [26]. Jiang, J., & Conrath, D. (1997). Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. *10th International Conference on Research in Computational Linguistics*, (pp. 19-33). Taipei.
- [27]. Jie, W., Stepane, B., & Beng, O. (2000). Mining Term Association Rules for Automatic Global Query Expansion: Methodology and Preliminary Results. *1st International Conference on Web Information Systems Engineering* (p. 366). Hong Kong: IEEE Computer Society.
- [28]. Koopman, B., Brunza, P., Sitbon, L., & Lawley, M. (2011). AHERC and QUT at TREC 2011 Medical Track: A Concept-based Information Retrieval Approach. *TREC*. Maryland.
- [29]. Leacock, C., & Chodorow, M. (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification*. (C. Fellbaum, Ed.) Cambridge: MIT Press.

- [30]. Limsopatham, N., Macdonal, C., & Ounis, I. (2013). A Task-specific Query and Document Representation for Medical Records Search. *Proceedings of the 35th European Conference on Advances in Information Retrieval* (pp. 747-751). Berlin: Springer-Verlag.
- [31]. Lin, D. (1998). An Information-Theoretic Definition of Similarity. *15th International Conference on Machine Learning* (pp. 296-304). Madison: Morgan Kaufmann.
- [32]. Liu, B., Hsu, W., & Ma, Y. (1998). Integrating Classification and Association Rule Mining. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 80-86). AAAI Press.
- [33]. Liu, Y., McInnes, B., Meaux, G., & Pakhomov, S. (2012). Semantic Relatedness Study using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and Wordnet. *ACM International Health Informatics Symposium* (pp. 363-372). Miami: ACM press.
- [34]. M.Voorhees, E., & Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. *TREC*.
- [35]. M.Voorhes, E., & Tong, R. (2011). Overview of the TREC 2011 Medical Records Track. *TREC*.
- [36]. Madria, S., Bhowmick, S., Ng, W., & Lim, E. (1999). Research Issues in Web Mining. *1st International Conference on Data Warehousing and Knowledge Discovery* (pp. 303-312). Springer.
- [37]. McCormick, T., Rudin, C., & Madiagan, D. (2012). Bayesian Hierarchical Rule Modeling for Predicting Medical Conditions. *The Annals of Applied Statistics* , 6 (2), 652-668.
- [38]. McInnes, B., Pedersen, T., & Pakhomov, S. (2009). UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. *American Medical Informatics Association Symposium*. San Fransisco.
- [39]. Ordonez, C. (2006). Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction. *IEEE Transactions on Information Technolog in Biomedicine* , 10 (2), 334-343.

- [40]. Ordonez, C. (2006). Comparing Association Rules and Decision Trees for Disease Prediction. *International Workshop on Healthcare Information and Knowledge Management* (pp. 17-24). Virginia: ACM.
- [41]. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A High Performance and Scalable Information retrieval Platform. *ACM SIGIR Workshop on Open Source Information Retrieval*. Seattle: ACM.
- [42]. Patwardhan, S., & Pederson, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. *EACL 2006 Workshop, Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, (pp. 1-8). Trento.
- [43]. Pederson, T., Pakhomov, S., Patwardhan, S., & Chute, C. (2007). Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *Journal of Biomedical Informatics*, 40 (3), 288-299.
- [44]. Peng, J., Macdonald, C., He, B., Plachouras, V., & Ounis, I. (2007). Incorporating Term Dependency in the DFR Framework. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 843-844). Amsterdam: ACM.
- [45]. Pluye, P., Grad, R., Dunikowski, L., & Stephenson, R. (2005). Impact of Clinical Information Retrieval Technology on Physicians: A Literature Review of Quantitative, Qualitative and Mixed Methods Studies. *International Journal of Medical Informatics*, 74 (9), 745-768.
- [46]. Rada, R., Mill, H., Bicknell, E., & Blettner, M. (1989). Development and Application of a Metric on Semantic Nets. *IEEE transactions on systems, man and cybernetics*, 19 (1), 17-30.
- [47]. Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 448-453). Montreal: Morgan Kaufmann.
- [48]. Robertson, S., & Jones, S. (1988). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27 (3), 129-146.
- [49]. Rocchio, J. (1971). Relevance Feedback in Information Retrieval. In G. Salton (Ed.), *the SMART Retrieval System: Experiments in Automatic Document Processing*, (pp. 313-323).

- [50]. Rohian, H., An, A., Zhao, J., & Huang, X. (2009). Discovering Temporal Associations among Significant Changes in Gene Expression. *IEEE International Conference on Bioinformatics and Biomedicine* (pp. 416-423). Washington: IEEE Computer Society.
- [51]. Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* , 24 (4), 35-43.
- [52]. Smucker, M., Allan, J., & Carterette, B. (2007). A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (pp. 623-632). New York: ACM.
- [53]. Trieschnig, D., Hiemstra, D., Jong, F., & Kraaij, W. (2010). A Cross-Lingual Framework for Monolingual Biomedical Information Retrieval. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 169-178). Toronto: ACM.
- [54]. Wang, W., & Wu, Y. (2010). Mining Association Rules in Medical Data Based on Concept Lattice. *8th World Congress on Intelligent Control and Automation*, (pp. 5171-5174). Jinan.
- [55]. Wang, W., Yang, J., & Yu, P. (2000). Efficient Mining of Weighted Association Rules. *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining* (pp. 270-274). Boston: ACM.
- [56]. Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133-138). Las Cruces: Association for Computational Linguistics.
- [57]. Xie, W., & Wu, J. (2010). Mining Positive and Negative Weighted Association Rules in Medical Records Without User-specified Weights Based on HITS Model. *3rd International Conference on Biomedical Engineering and Informatics*, (pp. 2325-2329). Yantai.
- [58]. Ye, Z., Huang, X., & Lin, H. (2011). Incorporating Rich features to Boost Information Retrieval Performance: A SVM-regression Based Re-ranking Approach. *Expert Systems with Applications* , 38 (6), 7569-7574.

- [59]. Yin, X., Huang, X., & Li, Z. (2010). Promoting Ranking Diversity for Biomedical Information Retrieval Using Wikipedia. *Advances in Information Retrieval, 32nd European Conference on IR Research* (pp. 495-507). Milton Keynes: Springer.
- [60]. Yin, X., Huang, X., & Li, Z. (2011). Mining and Modeling Linkage Information from Citation Context for Improving Biomedical Literature Retrieval. *Information Processing and Management: An International Journal (IPM)* , 47 (1), 53-67.
- [61]. Yin, X., Huang, X., Hu, Q., & Li, Z. (2009). Boosting Biomedical Information Retrieval Performance through Citation Graph: An Empirical Study. *Advances in Knowledge Discovery and Data Mining* (pp. 949-956). Bangkok: Springer.
- [62]. Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., & Butt, L. (2012). Exploiting Medical Hierarchies for Concept-based Information Retrieval. *Proceedings of the 17th Australasian Document Computing Symposium* (pp. 111-114). Dunedin: ACM.

A. IR System's Settings

In this chapter we present the details of our system's settings along with scripts and execution commands.

A.1 Data Pre-processing

To pre-process the collection we use Terrier's TermPipeline. TermPipeline is used to process each term in the collection. . TermPipeline can remove all the terms that should not be indexed by removing general stop words. It is also able to transform terms to their roots by stemming. The default setting for TermPipeline is porter stemming and general stopwords. This is specified in Terrier.properties file by "termpipelines=Stopwords,PorterStemmer" and "stopwords.filename=stopword-list.txt". We use the above settings to pre-process our data. We also added specific stop words that are very frequent in medical domain including patient, hospital, medical, room and diagnosis.

A.2 Building Index

To index our collection we use `TRECCollection` class. This class parses TREC formatted corpora. It provides sequential access to the documents in TREC collections. This is specified in `Terrier.properties` file by “`trec.collection.class=TRECCollection`”. Since we are using `TRECCollection` to index our dataset we should specify the tags that should be indexed. We specify the document tags by “`TrecDocTags.doctag=report`” and “`TrecDocTags.idtag=checksum`”. For proximity (dependence) models we should record term positions (block) in the index. To apply DFR-based dependence model we use block indexing by specifying “`block.indexing=true`”. We also set “`block.size=1`” to record the accuracy of one block. We configure Terrier to accept documents with larger docnos by configuring the property “`docno.byte.length`”; we set this property to 60. Figure 22 shows the commands to index a TREC collection assuming that the operating system is Linux.

```
#get terrier setup for using a trec collection
bin/trec_setup.sh /local/collections/TRECCollection/
#rebuild the collection.spec file correctly
find /local/collections/TRECCollection / -type f | sort |grep -v info >
etc/collection.spec
#index the collection
bin/trec_terrier.sh -i
```

Figure 22. Commands to index collection

A.3 Documents Retrieval

For processing the queries we should specify the tags of the topics file to be processed. We do that by setting the property “`TrecQueryTags.process`”. In our experiments we set this property to `SingleLineTRECQuery` class. This class reads queries, one per line. For query expansion we set the “`expansion.documents`” property to 3 and the

“expansion.terms” property to 10. This means that the query expansion component extracts the 10 most informative terms from the 3 top-returned documents as the expanded query terms. To set the parameter beta of the Rocchio's query expansion we should set the “rocchio.beta” property. For parameter beta we use the default value of 0.4.

Figure 23 shows the commands to retrieve documents.

```
#use this file for the topics
trec.topics=/local/collections2/Topics.txt >> etc/terrier.properties
#run the retrieval process
bin/trec_terrier.sh -r
#run topics again with query expansion
bin/trec_terrier.sh -r -q
#to enable the dependence model
bin/trec_terrier.sh -r -Dmatching.dsms=DFRDependenceScoreModifier
```

Figure 23. Command to retrieve documents

A.4 Terrier's Properties

In the terrier.properties file, properties are indicated in the format “name=value” and comments lines start with #. Figure 24 shows the properties we use to do data pre-processing, documents indexing and documents retrieval.

```
#directory names
terrier.home= Path to terrier
#default controls for query expansion
querying.postprocesses.order=QueryExpansion
querying.postprocesses.controls=qe:QueryExpansion
#default controls for the web-based interface. SimpleDecorate is the simplest metadata decorator. For more
control, see Decorate.
querying.postfilters.order=SimpleDecorate,SiteFilter,Scope
querying.postfilters.controls=decorate:SimpleDecorate,site:SiteFilter,scope:Scope
#default and allowed controls
querying.default.controls=start:0,end:999
querying.allowed.controls=c,scope,qe,qemodel,start,end,site,scope
#document tags specification
#for processing the contents of
#the documents, ignoring DOCHDR
```

```
TrecDocTags.doctag=report
TrecDocTags.idtag=checksum
#set to true if the tags can be of various case
TrecDocTags.casesensitive=false
#query tags specification
trec.topics.parser=SingleLineTRECQuery
#To set the parameter beta of the Rocchio's query expansion
rocchio.beta=0.4
trec.model=BM25
docno.byte.length=60
#stop-words file
stopwords.filename=stopword-list.txt
#the processing stages a term goes through
termpipelines=Stopwords,PorterStemmer
trec.topics=Path to Topics
```

Figure 24. Terrier's Properties

B. Conceptual Mapping Settings

To map the medical terms in our collection to UMLS Metathesaurus concepts we use MetaMap Java API. MetaMap tokenizes text into sections, sentences, phrases, terms, and words and maps the nouns of the text to the best matching UMLS concept(s). To use MetaMap Java API, full MetaMap installation and Java 1.6 SDK or greater are required. For more information on installing, downloading, instantiating and using MetaMap Java API see http://metamap.nlm.nih.gov/README_javaapi.html.

C. Mapping Reports to Visits

Figure 25 presents the Java program for mapping patients' reports to visits. "report-to-checksum-mapping-table.txt" is the report-to-checksum mapping table where each line contains report's number and the report checksum. "report-to-visit-mapping-table.txt" is the report-to-visit mapping table where each line contains report's checksum, report's type code, and the visit ID.

```
class mapping
{
    public static void main(String args[])
    {
        String[] mapping = null;
        String[] report = null;
        File file = new File("report-to-checksum-mapping-table.txt");
        BufferedInputStream bin = null;
        File file2 = new File("report-to-visit-mapping-table.txt");
        BufferedInputStream bin2 = null;
        try{
            FileInputStream fin = new FileInputStream(file);
            long length = file.length();
            if (length > Integer.MAX_VALUE) {
                System.out.print(" File is too large");
            }
            bin = new BufferedInputStream(fin);
            byte [] contents = new byte[(int)length];
            int bytesRead=0;
            String strFileContents;
```

```

while( (bytesRead = bin.read(contents)) != -1){
    strFileContents = new String(contents, 0, bytesRead);
    mapping= strFileContents.split("\\s+");
}
}
catch(FileNotFoundException e){
    System.out.println("File not found" + e);
}
catch(IOException ioe){
    System.out.println("Exception while reading the file "+ ioe);
}
finally{
    try{
        if(bin != null)
            bin.close();
    }
    catch(IOException ioe){
        System.out.println("Error while closing the stream :"+ ioe);
    }
}
try{
    FileInputStream fin2 = new FileInputStream(file2);
    long length = file2.length();
    if (length > Integer.MAX_VALUE) {
        System.out.print(" File is too large");
    }
    bin2 = new BufferedInputStream(fin2);
    byte [] contents = new byte[(int)length];
    int bytesRead=0;
    String strFileContents;
    while( (bytesRead = bin2.read(contents)) != -1){
        strFileContents = new String(contents, 0, bytesRead);
        report= strFileContents.split("\\s+");
    }
}
catch(FileNotFoundException e){
    System.out.println("File not found" + e);
}
catch(IOException ioe){
    System.out.println("Exception while reading the file "+ ioe);
}
finally{
    try {
        if (bin2 != null)
            bin2.close();
    }
}

```



```

catch(IOException ioe){
System.out.println("Error while closing the stream :"+ ioe);
}
}
try{
String temp= "";
int j=0;
int k=0;
int s;
for (int l=2; l< mapping.length-1; l=l+3) {
FileWriter myfile = new FileWriter ("visit"+k+".txt",true);
Writer output = new BufferedWriter(myfile);
for (int m=1; m< report.length-1 ; m=m+2){
if( mapping[l-2].equals(report[m])) {
s=Integer.parseInt(report[m-1]);
j=s;
break;
}
}
FileInputStream fstream = new FileInputStream ("report"+j+".xml");
DataInputStream in = new DataInputStream(fstream);
BufferedReader br = new BufferedReader(new InputStreamReader(in));
if (mapping[l].equals(temp)){
}
else {
output.append("<?xml version="+1.0+" encoding="+UTF-8+" standalone="+no+"?>");
output.append("\n");
output.append("<visit>");
output.append("\n");
output.append("<vid>"+mapping[l]+"</vid>");
output.append("\n");
}
DocumentBuilderFactory dbFactory = DocumentBuilderFactory.newInstance();
DocumentBuilder dBuilder = dbFactory.newDocumentBuilder();
Document doc = dBuilder.parse(in);
Element root =doc.getDocumentElement();
NodeList nList = doc.getElementsByTagName("report");
for (int r=0; r< nList.getLength(); r++){
Node node = nList.item(r);
if (node.getNodeType() == node.ELEMENT_NODE){
Element element= (Element) node;
NodeList nodelist= element.getElementsByTagName("report_text");
Element element1= (Element) nodelist.item(0);
NodeList fstNm=element1.getChildNodes();
output.append((fstNm.item(0)).getNodeValue());
}
}
}

```

```
}  
temp = mapping[l];  
if (mapping[l+3].equals(temp)){  
}  
else{  
output.append("</visit>");  
output.append("\n");  
k++;  
}  
br.close();  
output.close();  
}  
}  
catch (Exception e){  
System.err.println("Error: " + e.getMessage());  
}  
}  
}
```

Figure 25. Mapping reports to visits

D. Re-ranking Script

Figure 26 Presents the Perl program for re-ranking the initial search results based on the semantic query context.

```
#!/usr/bin/perl
$alpha=$ARGV[0];
print "alpha=$alpha\n";
#-----
$indexQueryFile="QueryContext.txt";
$indexDocFile="Index.txt";
$baseline= " Baseline.txt";
# query list
$listQueries="101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121
122 123 124 125 126 127 128 129 131 132 133 134 135";
my @listq = split(' ', $listQueries);
my @listd = split(' ', $listDoc);
$out=" Reranking_alpha$alpha.res";
open(OUT, ">$out") or die("Unable to open file $out");
open(QFILE, $indexQueryFile) or die("Unable to open file");
# read file into an array
@indexQuery = <QFILE>;
close(QFILE);
%docconcepts=();
%tf=();
%ndoc=();
open(DFILE,$indexDocFile) or die("Unable to open file");
foreach $line (<DFILE>) {
    @dataline=split(/ /, $line);
    $visid=$dataline[0];
    $concept=$dataline[2];
    $docconcepts{$visid}=$docconcepts{$visid}.$concept." ";
}
```

```

$tf{$concept.$visid}=$dataline[3];
$ndoc{$concept.$visid}=$dataline[5];
}
close(DFILE);
open(BFILE, $baseline) or die("Unable to open file");
# read file into an array
@indexBaseline = <BFILE>;
close(BFILE);
$Ntotaldoc=17267;
foreach my $q (@listq){
print "processing query $q *****\n";
# extract the concepts of the query
#-----
%QConW=();
$NormQ=0;
foreach $lineQ (@indexQuery){
#print "query line $lineQ";
chomp ($line);
@dataline=split(/ /, $lineQ);
$qcand=$dataline[0];
if($qcand eq $q){
$QConW{$dataline[1]}=$dataline[2];
$NormQ+=$dataline[2]**2;
#print "$q $dataline[1] $dataline[2]\n";
}
}
#print "NormQ $NormQ\n";
# read the collection document index file and extract the list of documents, then for each document, extract
the concepts of the document in a hash table
#-----
%newscore=();
%Normdoc=();
%baselinescore=();
$maxbaselinescore=0;
foreach $linebase (@indexBaseline){
chomp ($linebase);
@dataline=split(/ /, $linebase);
$qcand=$dataline[0];
$rank=$dataline[3];
$score=$dataline[4];
if($qcand eq $q && $rank<=1000){
$doc=$dataline[2];
$baselinescore{$doc}=$score;
#print "query cand $qcand doc $doc score $score\n";
if($maxbaselinescore<$score){
$maxbaselinescore=$score;
}
}
}

```

```

}
$conceptsdoclist=$docconcepts{$doc};
@conceptsDoc=split(/ /,$conceptsdoclist);
$concn=$doc{$concept.$doc};
#print "doc $doc conc $concept tf $conctf ndoc $concn\n";
if(exists($QConW{$concept}))){
print "concept $concept exists in query and doc $doc\n";
$newscore{$doc}+=$conctf*log($Ntotaldoc/$concn)*$QConW{$concept};
}
$Normdoc{$doc}+=$(conctf*log($Ntotaldoc/$concn)**2;
} # end for each concept in doc
}
} # end each line in baseline
$maxconScore=0;
while (($doc,$score) = each(%newscore)){
$newscore{$doc}=$newscore{$doc}/(sqrt($Normdoc{$doc})*sqrt($NormQ));
if($maxconScore<$newscore{$doc}){
$maxconScore=$newscore{$doc};
}
}
while (($doc,$score) = each(%newscore)){
$newscore{$doc}=$alpha*$newscore{$doc}/$maxconScore + (1-
$alpha)*$baselinescore{$doc}/$maxbaselinescore;
}
print "reranking ***** \n";
$r=0;
@sortdoc = sort ( { $newscore{$b} <=> $newscore{$a} } keys(%newscore) );
foreach $doc (@sortdoc) {
print OUT "$q Q0 $doc $r $newscore{$doc} rrCosine-caption\n" ;
$r++;
last if($r==1001)
}
}
foreach my $concept (@conceptsDoc){
$conctf=$tf{$concept.$doc};
close(OUT);

```

Figure 26. Re-ranking script

E. TOPKRULES Algorithm

Figure 27 presents the implementation of TOPKRULES algorithm by SPMF which is an open-source data mining platform written in Java. TOPKRULES is an algorithm for mining the TOP-K association rules with a pattern growth approach and several optimizations. This algorithm takes three inputs (1) a transaction database, (2) a parameter k representing the number of association rules to be discovered and (3) a parameter $minconf$ representing the minimum confidence that the association rules should have. The transaction database is a file where each line of the file contains set of items representing a transaction. To run the algorithm the following command should be executed: `java -jar spmf.jar TopKRules Transactions.txt output.txt 1000 60%`. Where “Transactions.txt” represents the transaction database, “1000” represent the parameter k and 60% represents the parameter $minconf$.

```
package ca.pfv.spmf.test;
import java.io.UnsupportedEncodingException;
import java.net.URL;
import ca.pfv.spmf.algorithms.associationrules.TopKRules_and_TNR.AlgoTopKRules;
import ca.pfv.spmf.algorithms.associationrules.TopKRules_and_TNR.Database;
public class MainTestTopKRules {
public static void main(String [] arg) throws Exception{
// Load database into memory
Database database = new Database();
```

```
database.loadFile(fileToPath("contextIGB.txt"));
int k = 10;
double minConf = 0.8; //
AlgoTopKRules algo = new AlgoTopKRules();
algo.runAlgorithm(k, minConf, database);
algo.printStats();
algo.writeResultToFile("C:\\Patterns\\result.txt"); // to save results to file
}
public static String fileToPath(String filename) throws UnsupportedEncodingException{
URL url = MainTestTopKRules.class.getResource(filename);
return java.net.URLDecoder.decode(url.getPath(),"UTF-8");
}
}
```

Figure 27. Java Implementation of TopKRules algorithm

F. Perl Program for Modeling Query Domain Ontology

To model query domain ontology we use UMLS::Interface module, which is a Perl interface to the UMLS. To model the query domain ontology we use getRelations function provided in the UMLS::Interface module. This function returns the relations of a concept. Input to this function is a concept (\$concept) and output is an array (\$array) that contains strings of relations. We pass the query concept as input to this function. To model the query domain ontology we also use getRelated function. This function returns a list of concepts (@concepts) related to a concept \$concept through a relation \$rel. Inputs to this function are (1) a concept (\$concept) and (2) a string containing a relation (\$rel). Output of this function is an array of related concepts. We pass the query concept and the array that contains strings of relations (output of getRelations function) as inputs to this function. We develop the query domain ontology using the output of this function.

```
use UMLS::Interface;
$configfile="myconfig.txt";
my $umls = UMLS::Interface->new({"config" => $configfile});
open (MYFILE, '>>query-domain-ontology.txt');
open (FILE, "query-concepts.txt") or die ("unable to open file");
@data=<FILE>;
foreach $concept (@data) {
  chomp($concept);
  my $Relations = $umls->getRelations($concept);
```



```
foreach my $relation (@{$Relations}) {  
  my $related_Concepts = $umls->getRelated($concept, $relation);  
  foreach my $related_concept (@{$related_Concepts}) {  
    print MYFILE "$related_concept\n";  
  }  
}  
}
```

Figure 28. Perl program for modeling query domain ontology

G. Perl Program for Calculating Semantic Relatedness

Figure 29 presents the program that returns a semantic similarity score between two concepts. This program takes as input a file where each line of the file contains pair of concepts in the following format: cui1 \diamond cui2. The input file should be specified after “--infile” option. The measure should be specified after “--measure” option. Following is an example of the usage: `query-umls-similarity-webinterface.pl --infile Inputfile.txt --measure vector`.

```
use Getopt::Long;
use URI::Escape;
use LWP;
open (MYFILE, '>>output.txt');
eval(GetOptions( "version", "help", "url=s", "measure=s", "sab=s", "rel=s", "infile=s", "debug")) or die
("Please check the above mentioned option(s).\n");
# if debug is defined
my $debug = 0;
if(defined $opt_debug) { $debug = 1; }
# if help is defined, print out help
if( defined $opt_help ) {
    $opt_help = 1;
    &showHelp();
    exit;
}
```

```

# if version is requested, show version
if( defined $opt_version ) {
    $opt_version = 1;
    &showVersion();
    exit;
}
# At least 2 terms should be given on the command line.
if( !(defined $opt_infile) and (scalar(@ARGV) < 2) ) {
    print STDERR "At least 2 terms or CUIs should be given on the \n";
    print STDERR "command line or use the --infile option\n";
    &minimalUsageNotes();
    exit;
}
# set browser handler
my $browser = LWP::UserAgent->new;
# initialize variables
my $rel    = "";
my $sab    = "";
my $reldef = "";
my $sabdef = "";
my $rmeasure = "";
my $smeasure = "";
my $button = "";
my $measure = "";
my $url    = "";
# check the input options
&checkOptions();
# set the options
&setOptions();
# get the concept pairs
my $input = &loadInput();
# get the similarity for each input pair
foreach my $pair (@{$input}) {
    my ($input1, $input2) = split/</>/, $pair;
    if($debug) {
        print STDERR "$input1 : $input2\n";
        print STDERR "SAB: $sab\n";
        print STDERR "REL: $rel\n";
        print STDERR "SABDEF: $sabdef\n";
        print STDERR "RELDEF: $reldef\n";
        print STDERR "BUTTON: $button\n";
        print STDERR "MEASURES: $rmeasure $smeasure\n";
    }
}
# query the web interface
my $page = &queryWebInterface($input1, $input2);
# extract the similarity information

```

```

my $output = &extractInformation($page);
# print output
print MYFILE "$output\n";
my $num = 2;
while($num--){
sleep(1);
}
}
}
sub extractInformation{
my $page = shift;
if($page =~ /View errors/) {
$page =~ /<input type="text" name="word1" id="word1in" value="(.*?)" \>/;
my $word1 = $1;
$page =~ /<input type="text" name="word2" id="word2in" value="(.*?)" \>/;
my $word2 = $1;
return "-1<>$word1<>$word2";
}
else {
$page =~ /<p class="results">The (similarity|relatedness) of (.*?) \(<a href="\#" onclick="showWindow
\(\u\ms\ _wps\.cgi\?wps=(C[0-9]+)(\.\*?')?, \'\)\); return false;\>(C[0-9]+)</a> \) and (.*?) \(<a href="\#"
onclick="showWindow \(\u\ms\ _wps\.cgi\?wps=(C[0-9]+)(\.\*?')?, \'\)\); return false;\>(C[0-9]+)</a> \)
using (.*?) \(.?') is (.*?)\.\</p>/;
my $word1 = $2;
my $cui1 = $3;
my $word2 = $6;
my $cui2 = $7;
my $score = $11;
#return "$score<>$word1($cui1)<>$word2($cui2)";
return "$cui1"." "$cui2"." "$score";
}
}
}
sub queryWebInterface{
my $i1 = shift;
my $i2 = shift;
if($debug) { print STDERR "In queryWebInterface($i1, $i2)\n"; }
my $qurl = "$url/cgi-
bin/u\ms\_similarity.cgi?word1=$i1&word2=$i2&sab=$sab&rel=$rel&similarity=$smeasure&button=$butto
n&sabdef=$sabdef&reldef=$reldef&relatedness=$rmeasure";
while(1) {
my $resp = $browser->get($qurl);
if ($resp->is_success) { }
else { print $resp->status_line, "$qurl \n"; }
my $webpage = $resp->content;
return $webpage;
}
}
}

```

```

sub loadInput {
if($debug) { print STDERR "In loadInput\n"; }
my @input_array = ();
# if file is defined get the terms or cuis from the input file
if(defined $opt_infile) {
if($debug) { print STDERR "FILE ($opt_infile) DEFINED\n"; }
open(FILE, $infile) || die "Could not open file: $infile\n";
my $linecounter = 1;
while(<FILE>) {
chomp;
if($ =~ /\s*$/) { next; }
if($ =~ ^<>/) {
# escape the ' character on input if it exists
if(! ($ =~ /\'/)) { $ =~ s'/\//g; }
push @input_array, $_;
}
else {
print STDERR "There is an error in the input file ($infile)\n";
print STDERR "one line $linecounter. The input is not in the\n";
print STDERR "correct format. Here is the input line:\n";
print STDERR "$_\n\n";
exit;
}
}
}
# otherwise get them from the command line
else {
if($debug) { print STDERR "Command Line terms/cuis defined\n"; }
my $i1 = shift @ARGV;
my $i2 = shift @ARGV;
# escape the ' character on input if it exists
if(! ($i1 =~ /\'/)) { $i1 =~ s'/\//g; }
if(! ($i2 =~ /\'/)) { $i2 =~ s'/\//g; }
if($debug) { print STDERR "INPUT: $i1 $i2\n"; }
my $input = "$i1 <$i2";
push @input_array, $input;
}
return \@input_array;
}
# checks the user input options
sub checkOptions {
if($debug) { print STDERR "In checkOptions\n"; }
if(defined $opt_measure) {
if(! ($opt_measure =~ /\b(path|wup|lch|cdist|nam|vector|res|lin|random|jcn|lesk)\b/)) {
print STDERR "The measure ($opt_measure) is not defined for\n";
print STDERR "the UMLS-Similarity package at this time.\n\n";
}
}
}

```

```

&minimalUsageNotes();
exit;
}
}
if(defined $opt_sab) {
if(defined $opt_measure && $opt_measure=~/lesk|vector/) {
if(! ($opt_sab=~/MSH|SNOMEDCT|UMLS_ALL/)) {
print STDERR "The --sab $opt_sab is currently not available through the web interface\n";
print STDERR "for the relatedness measure ($opt_measure)\n";
&minimalUsageNotes();
exit;
}
}
else {
if(! ($opt_sab=~/MSH|OMIM|SNOMEDCT|FMA/)) {
print STDERR "The --sab $opt_sab is currently not available through the web interface\n";
print STDERR "for the similarity measure (";
if(defined $opt_measure) { print STDERR "$opt_measure)\n"; }
else { print STDERR "path)\n"; }
}
&minimalUsageNotes();
exit;
}
}
}
if(defined $opt_rel) {
my @rels = split//, $opt_rel;
my $cui = 0; my $p = 0; my $c = 0; my $rb = 0 ; my $rn = 0;
foreach my $rel (@rels) {
if($rel=~/PAR/) { $p++; }
elsif($rel=~/CHD/) { $c++; }
elsif($rel=~/RN/) { $rn++; }
elsif($rel=~/RB/) { $rb++; }
elsif($rel=~/CUI/) { $cui++; }
else {
print STDERR "The relation $rel is not available in the webinterface.\n";
&minimalUsageNotes();
exit;
}
}
}
if(defined $opt_measure && $opt_measure=~/lesk|vector/) {
if(! (($cui > 0 && $p > 0 && $c > 0 && $rb > 0 && $rn > 0) ||
($cui > 0 && $p < 1 && $c < 1 && $rb < 1 && $rn < 1))) {
print STDERR "The --rel $opt_rel is currently not available through the web interface\n";
print STDERR "for the relatedness measure ($opt_measure).\n";
&minimalUsageNotes();
}
}

```

```

exit;
}
}
else {
if($cui > 0) {
print STDERR "The CUI relation is only available for the relatedness measures.\n\n";
&minimalUsageNotes();
exit;
}
if($p > 0 && $c > 0 && $rb > 0 && $rn > 0) {
print STDERR "The options PAR/CHD/RB/RN is currently not available\n";
&minimalUsageNotes();
exit;
}
if($p > 0 && $c < 1) {
print STDERR "Missing CHD option in PAR/CHD\n";
&minimalUsageNotes();
exit;
}
if($p < 1 && $c > 0) {
print STDERR "Missing PAR option in PAR/CHD\n";
&minimalUsageNotes();
exit;
}
if($rb > 0 && $rn < 1) {
print STDERR "Missing RN option in RB/RN\n";
&minimalUsageNotes();
exit;
}
if($rb < 1 && $rn > 0) {
print STDERR "Missing RB option in RB/RN\n";
&minimalUsageNotes();
exit;
}
}
}
}
}
# set user input and default options
sub setOptions {
if($debug) { print STDERR "In setOptions\n"; }
my $default = "";
my $set = "";
# set url
$url = "http://atlas.ahc.umn.edu/";
if(defined $opt_url) {
$url = $opt_url;
}
}
}
}
}

```

```

$set := " --set $url\n";
}
else { $default := " --default $url\n"; }
# set file
if(defined $opt_infile) {
$infile = $opt_infile;
$set := " --infile $opt_infile\n";
}
# set the measures
$measure = "path";
$measure = "vector";
$measure = "similarity";
if(defined $opt_measure) {
$set := " --measure $opt_measure\n";
if($opt_measure =~ /lesk|vector/) {
$button = "Compute+Relatedness";
$measure = $opt_measure;
$measure = "relatedness";
}
else {
$button = "Compute+Similarity";
$measure = $opt_measure;
}
}
else {
$button = "Compute+Similarity";
$default := " --measure $measure\n";
}
}
$measure = "MSH";
$measure = "UMLS_ALL";
if(defined $opt_sab) {
if($measure eq "similarity") { $sab = $opt_sab; }
else { $sabdef = $opt_sab; }
}
else {
if($measure eq "similarity") { $default := " --sab $sab\n"; }
else { $default := " --sab $sabdef\n"; }
}
}
# get the relation options for similarity and relatedness
$reldef = "CUI/PAR/CHD/RB/RN";
$rel = "PAR/CHD";
if(defined $opt_rel) {
my @rels = split~/, $opt_rel;
my $cui = 0; my $p = 0; my $c = 0; my $rb = 0; my $rn = 0;
foreach my $rel (@rels) {
if($rel =~ /PAR/) { $p++; }
}
}

```



```

if($rel==~/CHD/) { $c++; }
if($rel==~/RN/) { $rn++; }
if($rel==~/RB/) { $rb++; }
if($rel==~/CUI/) { $cui++; }
}
if($cui > 0 && $p > 0 && $c > 0 && $rb > 0 && $rn > 0) { $reldef = "CUI/PAR/CHD/RB%2fRN"; }
if($cui > 0 && $p < 1 && $c < 1 && $rb < 1 && $rn < 1) { $reldef = "CUI"; }
if($p > 0 && $c > 0) { $rel = "PAR/CHD"; }
if($rb > 0 && $rn > 0) { $rel = "RB/RN"; }
$set .= " --rel $opt_rel\n";
}
else {
if($measure eq "similarity") { $default .= " --rel PAR/CHD"; }
else { $default .= " --rel CUI/PAR/CHD/RB/RN"; }
}
# set the relation options for the web browser
$reldef==s~//%2F/g;
$rel==s~//%2F/g;
# check settings
if($default eq "") { $default = " No default settings\n"; }
if($set eq "") { $set = " No user defined settings\n"; }
# print options
print STDERR "Default Settings:\n";
print STDERR "$default\n";
print STDERR "User Settings:\n";
print STDERR "$set\n";
}

```

Figure 29. Perl program for calculating semantic relatedness of pair of concepts

H. Evaluation Scripts

To evaluate the results we use `trec_eval`. `Trec_eval` is the standard tool used by the TREC community for evaluating an ad hoc retrieval run, given the results file (`results.txt`) and a standard set of judged results (`qrels-trec.txt`). Figure 30 shows commands to evaluate the results.

```
# to evaluate the results
./trec_eval /local/qrels-trec.txt /var/results/results.txt
# to evaluate the results query by query
./trec eval -q /local/qrels-trec.txt /var/results/results.txt
```

Figure 30. Command to evaluate the results

I. Topics of the TREC Medical Records Track

Topics of the TREC medical records track 2011 are presented in Figure 31 (Note that topic 130 was dropped by TREC organizers because it has no relevant visits). Figure 32 presents conceptual representation of topics.

- 101 Patients with hearing loss
- 102 Patients with complicated GERD who receive endoscopy
- 103 Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
- 104 Patients diagnosed with localized prostate cancer and treated with robotic surgery
- 105 Patients with dementia
- 106 Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer
- 107 Patients with ductal carcinoma in situ (DCIS)
- 108 Patients treated for vascular claudication surgically
- 109 Women with osteopenia
- 110 Patients being discharged from the hospital on hemodialysis
- 111 Patients with chronic back pain who receive an intraspinal pain-medicine pump
- 112 Female patients with breast cancer with mastectomies during admission
- 113 Adult patients who received colonoscopies during admission which revealed adenocarcinoma
- 114 Adult patients discharged home with palliative care home hospice
- 115 Adult patients who are admitted with an asthma exacerbation
- 116 Patients who received methotrexate for cancer treatment while in the hospital
- 117 Patients with Post-traumatic Stress Disorder
- 118 Adults who received a coronary stent during an admission
- 119 Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes
- 120 Patients admitted for treatment of CHF exacerbation
- 121 Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and plavix
- 122 Patients who received total parenteral nutrition while in the hospital
- 123 Diabetic patients who received diabetic education in the hospital
- 124 Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma
- 125 Patients co-infected with Hepatitis C and HIV
- 126 Patients admitted with a diagnosis of multiple sclerosis
- 127 Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension
- 128 Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op
- 129 Patients admitted with chest pain and assessed with CT angiography
- 131 Patients who underwent minimally invasive abdominal surgery
- 132 Patients admitted for surgery of the cervical spine for fusion or discectomy
- 133 Patients admitted for care who take herbal products for osteoarthritis
- 134 Patients admitted with chronic seizure disorder to control seizure activity
- 135 Cancer patients with liver metastasis treated in the hospital who underwent a procedure

Figure 31. . Medical TREC 2011 Topics

101 C0030705 C1384666 C2029884
102 C0030705 C0231242 C0017168 C1514756 C1552424 C0014245
103 C0070668 C1522326 C2350012 C0014118 C0343401 C0014118 C1265292 C0014118
104 C0030705 C0011900 C0392752 C0376358 C0392752 C0600139 C1334407 C1278980 C0796563 C1278980 C1334407 C0033572 C0796563 C0033572 C1522326
C0336537 C0543467 C0336537 C0038894 C0336537 C0038895 C1705185 C0543467 C1705185 C0038894 C1705185 C0038895
105 C0030705 C0497327
106 C0030705 C0032743 C0024485 C1552358 C1552357 C0040405 C0332305 C1516647 C0150369 C0998265 C2707253 C0006826 C0027651 C1306459
107 C0030705 C1527349 C1176475
108 C0030705 C1522326 C1558950 C1456822 C0543467 C1801960 C1456822 C0543467
109 C0043210 C0029453
110 C0030705 C0012621 C0030685 C1510665 C0019994 C1524112 C0019004
111 C0030705 C0740418 C1514756 C1283188 C1555846 C1706421 C1283188 C1555846 C0182537 C1283188 C1555808 C1706421 C1283188 C1555808 C0182537
C1283188 C1555870 C1706421 C1283188 C1555870 C0182537
112 C0015788 C0030705 C1705498 C0030705 C1705497 C0030705 C0678222 C0006142 C0024881 C0809949
113 C0001675 C0030705 C1706450 C0030705 C1514756 C0009378 C0809949 C0443289 C0001418
114 C0001675 C0030705 C1706450 C0030705 C0012621 C0030685 C0442519 C0030231 C0442519 C0005555 C0442519 C1555983 C0442519 C0019947
115 C0001675 C0030705 C1706450 C0030705 C0809949 C1546484 C1547428 C0004096
116 C0030705 C1514756 C0025677 C0920425 C1510665 C0019994
117 C0030705 C0038436
118 C0001675 C1514756 C0687568 C0809949
119 C0001675 C0030705 C1706450 C0030705 C0449450 C0562508 C0860062 C0175668 C1522484 C0205436 C0011854
120 C0030705 C0809949 C1546484 C1547428 C1533734 C1705169 C0007111 C1522326 C0039798 C0018802
121 C0030705 C0010068 C0449450 C0562508 C2077962 C0948089 C1947971 C1998497 C0633084
122 C0030705 C1514756 C0030548 C2064719 C1510665 C0019994
123 C0241863 C0030705 C1514756 C0204935 C1510665 C0019994
124 C0030705 C0150312 C2003896 C0449450 C1510665 C0019994 C0332189 C0205178 C1517945 C0042789 C2707266 C0175668 C1522484 C0205436 C0017601
C0997768 C1962986
125 C0030705 C0439663 C0019196 C0220847 C0019682 C0019693 C0019699
126 C0030705 C0809949 C1546484 C1547428 C0011900 C1704656 C0944984 C0945731 C1704338 C1417325 C0026769
127 C0030705 C0809949 C1546484 C1547428 C0028756 C0277555 C0011847 C0011849 C0011860 C1963138 C0020538
128 C0030705 C0809949 C1546484 C1547428 C0022122 C0019552 C0187769 C1522326 C0009117 C2598133 C0009117 C0802604 C0009117 C0013227 C0687676
C1704687
129 C0030705 C0809949 C1546484 C1547428 C0008031 C1516048 C1536105
131 C0030705 C0205281 C0198482
132 C0030705 C0809949 C1546484 C1547428 C0543467 C0038894 C0038895 C0728985 C1269538 C0332466 C1293131 C0206078
133 C0030705 C0809949 C1546484 C1547428 C0580931 C1947933 C1515187 C0376667 C1514468 C0029408
134 C0030705 C0809949 C1546484 C1547428 C1555457 C0014544 C1547296 C0014544 C0205191 C0014544 C0008679 C1959629 C0008679 C0036572
C2587213 C1882979 C0243148 C1550141 C1148454
135 C1516213 C0494165 C1522326 C1510665 C0019994 C1546467 C0184661 C2700391 C0945766 C0944777 C1948041

Figure 32. Medical TREC 2011 Concept-based Topics

J. Sample Raw data

A sample of medical TREC 2011 report is presented in Figure 33. Figure 34 presents a sample visit that contains it's associated reports.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<checksum>20051127OP-cQsnkGImzZbN-848-71049104</checksum>
<subtype>ORTHO OP</subtype>
<type>OP</type>
<chief_complaint>LFT LEG PAIN</chief_complaint>
<admit_diagnosis>730.27</admit_diagnosis>
<discharge_diagnosis>250.81,707.14,403.91,428.0,711.06,276.7,424.1,416.0,730.27,250.51,362.01,414.8,2
44.0,272.4,</discharge_diagnosis>
<year>2007</year>
<downlaod_time>2009-10-05</downlaod_time>
<update_time/>
<deid>v.6.22.08.0</deid>
<report_text>[Report de-identified (Safe-harbor compliant) by De-ID v.6.22.08.0]
PROCEDURES:
TITLE OF OPERATION: IRRIGATION AND DEBRIDEMENT OF LEFT KNEE.
ANESTHESIA: General.
COMPLICATIONS: None.
PREOPERATIVE DIAGNOSIS(ES): SEPTIC ARTHRITIS, LEFT KNEE.
POSTOPERATIVE DIAGNOSIS(ES): SEPTIC ARTHRITIS, LEFT KNEE.
HISTORY AND INDICATIONS: The patient is a **AGE[in 60s]-year-old female with a history of end-
stage renal disease and hemodialysis with vasculopathy who by history, examination, and laboratory
studies had a septic arthritis of the left knee. Preoperatively, I spoke to the patient at great length. I spoke
to her and her daughter about the risks and benefits of surgical intervention. We talked about complications
of anesthesia, septic arthritis, continued pain, neurovascular surgery, need for future surgeries, soft tissue
complications etc. I explained to that irrigation and debridement of septic arthritis is indicated and we
talked about this at great length. After thorough a discussion about the risks and benefits of surgery, the
patient gave informed consent.
```

DESCRIPTION OF OPERATION: The patient was identified as the patient. She was taken to the operating room where she was placed supine on a table. Anesthesia had attempted to place a block; however, this did not work and therefore she needed to be intubated. After successful intubation, a nonsterile tourniquet was carefully placed high in the left thigh. The left leg was then prepped and draped in the usual sterile fashion while making sure to isolate the left foot on which she had surgery a few days prior. The leg was elevated for 120 seconds and then the tourniquet was inflated. A small approximately 5 cm parapatellar arthrotomy was performed sharply with a knife. This was taken down into the joint sharply. Immediately significant amount of cloudy-looking fluid came out of the knee. This was sent for culture. After evacuating the fluid, the knee was pulse irrigated with 3 L of solution. After this, we reexamined the knee. There was no further sign of purulence. The skin bleeders were coagulated. Again, 3 more liters of pulse irrigation were used to clean out the knee. After successfully accomplishing this, the arthrotomy was closed with 0 Vicryl in a watertight fashion. The skin was then closed carefully with interrupted 3-0 nylon sutures. A sterile consisting of Xeroform, 4x4's, Webril, and Ace wrap were applied. The patient was awakened from anesthesia. Earlier the tourniquet had been deflated prior to closure. There were no complications during this procedure. She was brought to the PACU in a stable condition.

</report_text>
</report>

Figure 33. Sample of medical TREC 2011 report

```
<?xml version=1.0 encoding=UTF-8 standalone=no?>
<vid>+K3ygdYUBVj</vid>
<visit>
Report de-identified (Safe-harbor compliant) by De-ID v.6.22.07.0]
Mr. **NAME[AAA] is a distant smoker who is now admitted with C. diff. In early **DATE[Oct], he was admitted with a pneumonia and new stroke. It was feared at that time that he may have had an endocarditis with an embolic stroke. CT scan performed in early **DATE[Oct] showed a severe right upper lobe dense consolidated pneumonia with air bronchograms which were not consistent with postobstructive pneumonia. He was treated with Zosyn, after results of his sputum culture came back revealing Pseudomonas cepacia. Apparently, he had had a bronchoscopy done at **NAME[VVV] prior to that admission that grew some yeast, and there is concern about a possible yeast infection. He was transferred to rehabilitation on **DATE[Oct 9 2007], on Zosyn 4.5 IV q.12 hours for 2 weeks total duration. He developed diarrhea yesterday, and therefore was readmitted with possible C. diff colitis.
PAST MEDICAL HISTORY: Remarkable for known coronary disease status post bypass grafting. He is a very distant smoker. No history of obstructive lung disease. He had a history of rectal carcinoma status post a partial colectomy with a colostomy.
SOCIAL HISTORY: Unhelpful.
REVIEW OF SYSTEMS: Pertinent for cough productive of frothy white sputum. No hemoptysis. Denies fevers, chills, or sweats. No chest pain, PND, orthopnea, or pedal edema. He has been on oxygen since he was admitted to the hospital in **DATE[Oct] with pneumonia.
PHYSICAL EXAMINATION: He is afebrile with a blood pressure of 130/68, his heart rate is in the mid 90s, respiratory rate is 18. His saturations in the mid 90s on 3 L nasal cannula. He is an awake, alert, thin, ill-appearing gentleman in no distress. Neck is supple without JVD. His lungs are bronchial on the right, scattered rhonchi, no wheezes. Regular rate and rhythm on cardiac exam, slightly tachycardic, no S3 or rub. Abdomen is soft, scaphoid, no organomegaly or masses. Ostomy in the left lower quadrant with a bag filled with greenish liquid stool. He had no significant edema.
```

LABORATORY DATA: Sodium 136; potassium 3.4; chloride 106; bicarbonate 21; BUN and creatinine of 29 and 2.4; blood glucose of 97; white count of 17,300 down from 19,800; H and H are 9.9/28.9 with 265,000 platelets. Sputum is growing rare normal flora. His CT scan shows persistence, although improved aeration in the right upper lobe. The left lower lobe and lateral lingula show improvement in infiltrates. Mr. **NAME[AAA] is a gentleman that had a severe pneumonia. Pneumococcus was certainly a possible etiology that did not grow, although Pseudomonas cepacia did eventually grow and he was treated with Zosyn. Unfortunately, he seems to have developed C. diff, although those results are pending. He is currently on Flagyl for that. From a pulmonary standpoint, he still has significant radiographic abnormalities, but they have improved. The air bronchograms are inconsistent with a postobstructive pneumonia. At this time, he has received adequate antibiotics and I would withhold further antibiotics from a pulmonary standpoint. Mucomyst and nebulized bronchodilators are reasonable to help him clear his secretions, and he will need a followup CT scan in about 2 months to document further clearing of that upper lobe infiltrate.

Thank you very much.

REASON FOR ADMISSION: C. difficile colitis.

HISTORY OF PRESENT ILLNESS: Mr. **NAME[AAA] is an **AGE[in 80s]-year-old gentleman with a recently complicated past medical history. He has a past history of coronary artery disease status post CABG in 2001. He has hypertension and he has a history of colorectal cancer for which he underwent surgery and has a colostomy in place. This was done 15 years ago. In late **DATE[Sep] the patient sustained a left-sided clavicular fracture and shoulder injury. For this he was admitted at **INSTITUTION. He was being managed there and subsequently was transferred to **INSTITUTION fro sepsis. He was admitted at the ICU at **INSTITUTION where he was managed for his sepsis. Blood cultures from the outside hospital showed E. coli. The patient was managed with IV antibiotics. During his stay in the ICU he developed a left-sided parietal infarction. Infective endocarditis was ruled out on a transthoracic echocardiogram. The patient was treated with Aggrenox. Also, during this time period the patient had shortness of breath and a cough productive of frothy white sputum. His sputum cultures grew (_____). This was thought to be a contamination and a CT scan of the chest showed multilobar pneumonia. The patient was treated with IV Zosyn. He was kept on 1 liter of oxygen by nasal cannula, flutter valve

treatment, physical therapy, DuoNeb, and expectorant treatment was given. Also, during the hospital stay the patient was markedly depressed and was kept on Zoloft. The patient was transferred on **DATE[Oct 9 2007], to

**INSTITUTION Rehabilitation at **PLACE where his antibiotics for 2 weeks for a completion of a 2-weeks' course were continued. About a week ago the patient was transferred to **NAME[VVV] again with productive cough of frothy sputum and shortness of breath. Over there, Dr. (_____) did a bronchoscopy. According to his note, there was no endobronchial lesion in the right upper lobe. In the post bronchoscopy sputum and washings the patient grew fungus. By this time on evaluating his imaging, he mentions in his note that originally there was a right lower lobe, right upper lobe, and left lower lobe infiltrate which are resolving. The right upper lobe infiltrate, however, is still persistent. The patient was discharged back to Southside Rehabilitation on IV Diflucan, starting on **DATE[Oct 27 2007]. At Southside Rehabilitation the patient continued to have his productive cough and yesterday IV Zosyn was added by the patient's PCP. He also added Flagyl 500 mg p.o. t.i.d. to cover for C. difficile colitis. Yesterday, the patient had a sudden increase in colostomy output. During 1 shift he had 1700 cc put out. Also, he was noted to have a fever. His white cell count yesterday had jumped up to 24,000. On **DATE[Oct 28 2007], it was 7.9. Neutrophil count was 72% and bandemia was 23%. A sepsis workup was done and the patient was positive for C. difficile colitis. Also, his UA was consistent with a UTI.

Today, the patient was noted to have a blood pressure of 80 to 90 systolic, was tachycardic, and was noted to have PACs. For this reason the patient was transferred to **INSTITUTION for further management.

PAST MEDICAL HISTORY: The patient had an MI in 1991. In 2001 he had a 4-vessel CABG. He has hypertension. Fifteen years ago he was diagnosed with rectal cancer. He underwent a surgery and sustained a colostomy from that surgery. Since then he has not had any chemotherapy or radiation therapy. He is not being followed up by any oncologist and neither does he have a cardiologist.

REVIEW OF SYSTEMS: As per his history of present illness. Also, the patient has anorexia, decreased p.o. intake, and nausea. He has been vomiting. He has intermittent fevers and chills. His fevers are up to 101. He complains

of shortness of breath on exertion; however, these days he is not exerting himself too much. As mentioned above, he has a cough productive of foamy white sputum. He has had significant weight loss; however, he does not know how much. He does complain of intermittent urinary incontinence.

MEDICATIONS:

1. Aggrenox 25/200 mg tablet 1 tablet b.i.d.
2. Lomotil 1 tablet q.6 h.
3. Diflucan 100 mg IV q.12 .
4. Humibid 2 tablets p.o. b.i.d.
5. Subcu heparin.
6. Ferrex 150 mg p.o. b.i.d.
7. Megace 40 mg p.o. b.i.d.
8. Metronidazole which was started yesterday, 500 mg p.o. t.i.d.
9. Multivitamins.
10. Nystatin local application.
11. Protonix 40 mg p.o. daily.
12. Zosyn 4.5 mg IV q.8 h. which was started yesterday.
13. Probenecid 500 mg p.o. b.i.d.
14. Triamcinolone local application b.i.d.
15. Effexor 37.5 mg p.o. daily.
16. Verapamil 90 mg p.o. daily.
17. Flagyl was entered twice 500 mg q.6 h. IV.
18. Tylenol.
19. Zofran.
20. Percocet.
21. Benazepril 20 mg p.o. daily.

ALLERGIES: THE PATIENT HAS NO KNOWN DRUG ALLERGIES.

FAMILY HISTORY: Noncontributory.

SOCIAL HISTORY: The patient used to live alone up until **DATE[Sep]. Prior to that, especially in the summers, he did extensive gardening work. He is a retired coal miner. He has a remote history of tobacco use. He stopped about 50 years ago and has an approximate 10-year-pack history. He has occasional ethanol intake. He denies IV drug use.

PHYSICAL EXAMINATION: This is an elderly gentleman. He looks very wasted and cachectic. He is alert and cooperative. Blood pressure 112/52, pulse 100, respiratory rate 18, temperature 36.7, saturations 95. **HEENT:** The patient is very cachectic and he has temporal wasting. Otherwise, extraocular muscles are intact. Pupils are bilaterally equally reactive to light. He has moist mucous membranes. No cervical lymphadenopathy. There is no JVP. **Cardiovascular Exam:** There is a central sternotomy scar. He has a normal S1, S2 with a systolic murmur. He is slightly tachycardic. **Pulmonary Exam:** The patient has

coarse crackles, right more than the left side. The crackles are in the lower and mid zone. There is mid and upper zone bronchial breathing on the right-hand side. On the left side there are coarse crackles, particularly at the bases. There is intermittent wheezing. Abdomen is soft, looks slightly distended, and he is tender diffusely. The left half of his abdomen - he has a colostomy in place with greenish feces. Gut sounds are audible but reduced. Lower Extremities: There is no edema. The patient has right upper and lower extremity weakness.

LABS: All labs are pending at this point in time.

ASSESSMENT AND PLAN: This is an **AGE[in 80s]-year-old gentleman with a complicated past medical history since the past 2 months or so. He presents from Southside Rehabilitation with the following issues.

1. Clostridium difficile colitis. The patient has been started on Flagyl 500 mg p.o. t.i.d. His Zosyn has been put on hold. We are also starting Lactinex 1 pack p.o. q.i.d. I will check an abdominal x-ray as the patient is complaining of abdominal tenderness. This is to rule out any toxic megacolon or perforation.

2. Multilobar pneumonia. The patient has already received Zosyn last month for 2 weeks for this. He has recently been started on Diflucan based on fungus that was grown in his bronchial washings. I do not have the details available of the bronchial report or the culture at this point in time. We will continue the IV Diflucan. We will check a CT scan of the chest without contrast. We will consult Dr. **NAME[UUU]'s group. We will also continue the patient's DuoNeb, Mucomyst, and guaifenesin and send off sputum cultures.

3. Dehydration and prerenal azotemia. Reportedly, the patient has had a bump in his creatinine. A couple of days ago, his creatinine was 0.9 and now it has increased to 3 at the rehabilitation. This most likely seems prerenal because of the dehydration and decreased p.o. intake as well as vomiting. We will generously hydrate the patient. The patient has been started on normal saline at 200 mL per hour. We will generously hydrate as the patient's ejection fraction is 60%. He does have a history of diastolic dysfunction. Therefore, we will be careful not to cause volume overload. Check urine electrolytes but we will hold off on any further renal workup. We will hold any nephrotoxic medications and renally adjust his medications.

4. Ectopic beats. I will check an EKG and place the patient on a monitor.

5. Deep vein thrombosis prophylaxis, gastrointestinal prophylaxis, PT, and OT consult. We will consult social work. The patient has been discussed with the attending.

CLINICAL HISTORY: Multilobar pneumonia. Evaluation for response.

COMPARISON: Previous chest CT scan from **DATE[Oct 04 2007].

TECHNIQUE: Helical CT imaging of the chest was obtained without intravenous or oral contrast with contiguous 5.0mm thick axial reconstructions using both lung and standard reconstruction algorithms.

FINDINGS: The areas of patchy airspace consolidation in the left upper lobe have resolved. There has been near complete resolution of the left lower lobe pneumonia. The left pleural effusion has resolved. There is persistent extensive dense consolidation in the right upper lobe which has decreased. The areas of airspace consolidation in the right lower lobe and right middle lobe have decreased.

IMPRESSION:

1. PERSISTENT VERY EXTENSIVE CONSOLIDATION IN THE RIGHT UPPER LOBE (WITH NORMAL PATENT AIRWAYS). THE DEGREE AND EXTENT OF RIGHT UPPER LOBE CONSOLIDATION HAVE DEFINITELY DECREASED FROM THE PREVIOUS STUDY OF **DATE[Oct 04 2007] CONSISTENT WITH RESOLVING PNEUMONIA. THERE ARE NO FINDINGS TO INDICATE ENDOBRONCHIAL OBSTRUCTING LESIONS.

2. INTERVAL DECREASE IN AREAS OF AIRSPACE CONSOLIDATION AND RIGHT LOWER LOBE AND RIGHT MIDDLE LOBE.

3. NEAR COMPLETE RESOLUTION OF AREAS OF CONSOLIDATION IN LEFT

UPPER LOBE AND LEFT LOWER LOBE.

4. RESOLVED LEFT PLEURAL EFFUSION.

END OF IMPRESSION.

Physical Medicine and Rehabilitation Consult.

I was asked by Dr. **NAME[ZZZ] to determine and comment on **NAME[BBB AAA]'s rehabilitation options and needs. Mr. **NAME[AAA] is an **AGE[in 80s]-year-old right-handed male admitted to **INSTITUTION on **DATE[Nov 1 2007]. The patient's chief complaint is increasing colostomy output and weakness. The patient's history began late **DATE[Sep 2007] when he fell incurring a clavicle fracture. He was treated at **INSTITUTION. While there, he developed sepsis and transferred to **INSTITUTION for further treatment and management. Blood cultures from **INSTITUTION grew E. coli. While at **INSTITUTION, the patient was found to have a right lung pneumonia on chest x-ray as well as CT scan. He completed a course Zosyn. During his stay, the patient developed a sudden onset right hemiparesis from a left parietal and bilateral inferior cerebellar CVA. The patient was placed on aspirin and Plavix. He was discharged to **INSTITUTION, **PLACE for a rehabilitation program. The patient stayed approximately a week and a half. He was readmitted to **INSTITUTION with a productive cough. Workup included bronchoscopy which revealed a fungal infection on cultures. The patient was placed on IV Diflucan and subsequently discharged back to **INSTITUTION, **PLACE.

Recently, the patient was admitted to **INSTITUTION with increasing colostomy output. He was noted to have leukocytosis with a white count of 24,000 and 23% bandemia. The patient was transferred to **INSTITUTION on **DATE[Nov 1 2007], for further treatment.

The patient was diagnosed as having C. difficile colitis. His C. difficile toxin assay was positive from his **DATE[Nov 1 2007], stool sample. His blood cultures have shown no growth. The patient has been placed on Lactobacillus along with Flagyl. He has shown clinical improvement with decreased colostomy output. His WBC count has decreased to 8.6. The patient was also dehydrated, and his creatinine since improved to 1.4. His chest CT showed persistent right upper lobe consolidation, but overall improved from previous radiographs.

From a functional standpoint, the patient requires minimal assistance with his transfer, minimal assistance to ambulate 5 feet using a wheeled walker, and requiring assistance to open his containers during meals.

PAST MEDICAL HISTORY: Coronary artery disease status post coronary artery bypass graft in 2001, history of MI, hypertension, colorectal CA status post colostomy 15 years ago, and hyperlipidemia.

FAMILY HISTORY: Positive for coronary artery disease.

ALLERGIES: THE PATIENT HAS NO KNOWN DRUG ALLERGIES.

MEDICATIONS: As listed in the 36-hour report. I have reviewed this list.

PERSONAL AND SOCIAL HISTORY: The patient is a retired coal miner. He lived alone prior to his admission. A former smoker. Drinks alcohol occasionally. He was independent prior to admission.

REVIEW OF SYSTEMS: The patient feels tired and weak. His right is weaker compared to his left. He denies any chest pain. He does have exertional dyspnea, but overall better. He reports cough, but this is also improved.

He denies any swallowing problems. Currently has a Foley catheter in place. He reports being continent prior to admission. He has no neck or back pain. He denies any numbness or paresthesias. Rest of his review of systems is

negative. PHYSICAL EXAMINATION: The patient is thin elderly male. He is awake. He is alert. He is oriented x3 in no acute distress. His voice is wet and gurgly, hypophonic, but easily understandable. Heart: Distant. Lungs: With scattered rhonchi. Abdomen: Soft and nontender. Extremities: He has trace bilateral pretibial edema. Skin: Dry and scaly. I did not look at his sacral coccyx for any pressure ulcers. Neurologic: He has full passive range of motion of both upper and lower extremities with minimal pain at

the end range. Manual testing, left upper extremity is 5/5, right upper extremity is 3 to 4 over 5; in both lower extremities, it is limited secondary to his position, but left TA and EHL is 5/5 and right TA and EHL is 4+/5. Muscle stretch reflexes are slightly brisker on the right. He does not have any clonus, and his plantar reflex is equivocal. He does have dysmetria in both upper extremities, more significant on the right. IMPRESSION: The patient is an **AGE[in 80s]-year-old male with a history of left parietal cerebrovascular accident and right hemiparesis. He is admitted with Clostridium difficile colitis. He has had extensive and complicated course since his initial admission in **DATE[Sep 2007]. He is not at his functional baseline with decline in functional mobility, self care, as well as speech.

RECOMMENDATIONS:

1. Consider modified barium swallow to evaluate his current swallowing status.
2. Agree with deep venous thrombosis prophylaxis.
3. The patient will require an admission to the acute inpatient rehabilitation service. He has the comprehensive therapy needs with physical therapy, occupational therapy, and speech and language pathology to address his functional decline as well as extensive active medical issues requiring close supervision and judicious management.

Thank you for allowing me to participate in his care.

FINDINGS:

Subacute near anatomically aligned extra-articular left distal clavicular fracture is seen with a small amount of callus. Bones are osteopenic.

IMPRESSION:

SUBACUTE NEAR ANATOMICALLY ALIGNED LEFT DISTAL CLAVICULAR FRACTURE WITH SMALL AMOUNT OF CALLUS.

END OF IMPRESSION.

FINAL DIAGNOSES:

1. Acute Clostridium difficile colitis.
2. Diarrhea secondary to the above.
3. Dehydration.
4. Hypertension, improved.
5. Acute renal failure, prerenal in nature, improved.
6. Questionable deep venous thrombosis in the right arm, but venous Dopplers are negative.
7. Anemia.
8. Recent pneumonia with persistent cough and sputum.
9. Severe cough and pulmonary congestion.
10. Rule out congestive heart failure, but doubt.
11. Recent cerebrovascular accident.
12. Nausea and intolerance to some medications.
13. Cardiac ectopies with sinus tachycardia and premature atrial contractions.
14. Recent cerebrovascular accident in early with right-side residual weakness.
15. Recent urosepsis with Escherichia coli in early **DATE[Oct], improved.
16. Recent severe acute renal failure, improved on this admission, and he came back with renal failure and improved again.
17. Depression.
18. Physical and mental deconditioning.
19. History of colon cancer, status post colectomy and colostomy.
20. History of coronary artery disease.
21. Previous myocardial infarction about 15 years ago.

22. Status post coronary artery bypass grafting about 5 or 6 years ago.

DISCHARGE MEDICATIONS:

1. Percocet 5 mg q.6h. as needed.
2. Mucomyst via nebulizer twice a day.
3. DuoNeb q.i.d.
4. Aggrenox 25/200 twice a day.
5. Xenaderm for certain decubitus areas.
6. Aranesp 40 mcg once a week.
7. Lovenox 45 mg b.i.d. stopped at the time of transfer.
8. Mucinex 1200 mg twice a day.
9. Heparin 5000 units subcutaneously q. 12h.
10. Lactinex one packet b.i.d. for one week.
11. Megace 40 mg p.o. b.i.d.
12. Flagyl 500 mg three times a day for one week.
13. Nasonex spray each nostril twice a day.
14. Protonix 40 mg daily.
15. Lotensin 20 mg p.o. daily.

HISTORY: Mr. **NAME[AAA] is an **AGE[in 80s]-year-old pleasant man who was discharged recently to **INSTITUTION Rehab after a long admission for throat, renal failure, sepsis, and severe pneumonia. He was doing well at the rehab; however, on the night before the day of admission, he started to have diarrhea. He did not respond to treatment. C. diff toxin was sent, but the patient became hypotensive, his creatinine went up to 2 and BUN became elevated. He became tachycardic, and showing irregular heart beats. For all these reasons, he was transferred to **INSTITUTION for further management and evaluation.

HOSPITAL COURSE: The patient was admitted with initial problems that included severe diarrhea confirmed later to be from C. diff colitis, it was treated with Flagyl and Lactinex and seemed to improve slowly. His other issue was hypotension and tachycardia related to dehydration and low volume. It was corrected with IV fluids. He had renal failure on admission, the creatinine was 2.0 and elevated BUN. That seemed to correct with IV fluids, creatinine improved down to 1. He was tachycardic and showing irregular heart beats on admission, that seemed to improve to normal sinus rhythm with controlled rate at the time of discharge. The patient had swelling in the right arm, there was question of DVT. However, venous Dopplers came back negative. The patient continues to have the problem with dysphagia. He is on dysphagia diet. However, his main complaint and problem seem to be persistent congestion in the lungs and sputum production and cough. He had recent pneumonia, which seemed to be getting better on CT scan and chest x-ray. I wonder if he has either components of microaspiration or postnasal drip which could be contributing. Appropriate treatments are being given to that. Other agents could be added later in the future. As the patient seemed to improve from the diarrhea standpoint and as his kidney function coming back to normal, his blood pressure and heart rate are improving, we felt that he could go back to the **INSTITUTION Rehab to continue physical therapy and rehabilitation. He will continue his medications.

LABORATORY DATA: Prior to discharge, his stool for occult blood was negative. His sodium was 140, potassium 3.9, chloride 119, BUN 17, creatinine 1.4, calcium 8.2. His INR was 21, TIBC 1 was 7 and iron saturation was 20%, ferritin 251.

Hemoglobin was 9.0, hematocrit 26.8, and platelet count 255.

</visit>

Figure 34. Sample of medical TREC 2011 visit