Arab American University

Faculty of Graduate Studies

QUERY REFORMULATION IN THE CONTEXT OF MEDICAL INFORMATION RETRIEVAL

By

Isra' A. Noor

Main Supervisor

Dr. Mohammed A. M. Maree

Co-Supervisor

Dr. Khaled S. Rabayah

This thesis was submitted

in partial fulfillment of the requirements for

the Master`s degree in Computer Science

May 2018

Query Reformulation in the Context of Medical Information Retrieval

By

Isra' A. Noor

This thesis was defended successfully on …………………and approved by:

Committee members	Signature
1. Supervisor Name: Dr. Mohammed Maree	....
2. Co- Supervisor Name: Dr. Khaled Rabayah	....
3. Internal Examiner Name: Dr. Mohammed Awad	....
4. External Examiner Name: Dr. Labib Arafeh	....

Declaration

This is to declare that the thesis entitled "Query Reformulation in The Context of Medical Information Retrieval" under the supervision of Dr. Mohammed A. M. Maree and Dr. Khaled S. H. Rabayah is my own work and does not contain any unacknowledged work or material previously published or written by another person, except where due reference is made in the text of the document.

Date: 10/6/2018

Name: Isra' Noor Signature:

Acknowledgements

I would like to thank my supervisor, Dr. Mohammed Maree, for the patient guidance, encouragement, and advice he has provided throughout my master thesis. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I would also like to thank my co-supervisor Dr. Khaled Rabayah for his support and engagement all the time.

I'd like to express my gratitude to the people who mean a lot to me, my father and my mother, for believing in me, letting me choose what I desired, and supporting me all the time. I thank you all for your selfless love, care, pain, and sacrifice to shape my life. I would never be able to pay back the love and affection you give to me. I also express my thankfulness to my brothers and sisters for their support and encouragement throughout my life.

Finally, I will never forget to thank my husband Malek, for his continued support and understanding during my pursuit of a master degree that made the completion of thesis possible. I greatly value his contribution and deeply appreciate his belief in me.

Thanks for all your encouragement

Abstract

In the recent years, search engines on the Web have become indispensable tools that assist users in finding information in almost any domain of interest. These search engines can be classified into two main broad categories; generic (known as horizontal) and domain-specific (known as vertical) search engines. Both types of search engines exploit different algorithms and techniques to first index documents on the Web and accordingly retrieve the most relevant results that satisfy users' information needs. In this research work, our focus has been on the second category of search engines i.e. domain-specific search, namely in the medical domain. When using medical search engines, the precision of the retrieved results is mainly dependent on two major factors. First, users (laypeople or medical professionals) need to submit vocabularies that best describe their information needs. Second, the quality of the returned results is largely based upon the effectiveness of the employed techniques and medical knowledge resources used by such search engines.

Several systems and approaches have been proposed to address problems associated with each of these factors independently; such as using general terms to express laypeople needs instead of using professional medical concepts, and using knowledge resources with limited coverage of medical concepts which lead to a semantic gap between user's queries and their related medical documents. However, little attention has been paid to jointly address problems in both factors. In this research, we have planned to investigate the impact of exploiting medical semantic resources and information retrieval techniques on 1) reformulating the given users' queries through enriching them with semantically related terms and 2) indexing documents in the medical domain to improve quality of the retrieved search results. To demonstrate the effectiveness of the proposed system, we have conducted several experiments using the Conference and Labs Evaluation Forum (CLEF) eHealth2014 dataset. Additionally, we have evaluated the effectiveness of the proposed system (MedIR) against three state-of-the-art medical information retrieval systems - that used the same dataset for system evaluation - based on the well-known precision indicator. The conducted experiments using the above dataset show that the proposed approach has achieved a precision of 0.7945, while the precision of the other similar medical information retrieval systems we compared with was (0.7560, 0.7540, and 0.7400). The previous precision values demonstrate that the quality of the retrieved results of our approach has improved by around five to seven percent compared to other similar systems.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

MIR – Medical Information Retrieval

NLP – Natural Language Processing

UMLS – Unified Medical Language System

MeSH – Medical Subject Headings

SNOMED-CT – Systematized Nomenclature of Medicine - Clinical Terms

ICD-CM – International Classification of Diseases- Clinical Modification

CLEF – Conference and Labs Evaluation Forum.

HON – Health On the Net

VSM – Vector Space Model

Tf–Idf - term frequency–inverse document frequency

PRF - Pseudo Relevance Feedback

JSF – Java Serve Faces

IR - Information Retrieval

QR - Query Reformulation

CSS - Cascading Style Sheets

HTML - Hypertext Markup Language

API - Application Programming Interface

DOM - Document Object Model

Chapter 1 - Introduction

1.1 Background

A wide variety of users, including patients, researchers, general physicians, and specialists with specific expertise such as radiologists or oncology specialists are interested in medical information. According to the Pew Research Center report from 2013, of the 85% of U.S. adults that uses the Internet 72% have looked for health related information within the past year (Fox and Duggan, 2013). In this way, the Internet has become one of the most common sources of medical information. To access this information, users need to have intelligent search methods and techniques that enable them to obtain valuable results that accurately fulfill their needs (Zickuhr, Susannah Fox, 2013). The diversity of users, the information they need, and their knowledge level of medical data have a large impact on the effectiveness of medical information retrieval (MIR) systems (Soldaini et al., 2016, Stanton et al., 2014). The diversity of the knowledge level of the user affects the way in which the query is formulated, and also affects the type and quality of the retrieved results (Palotti et al., 2014, White et al., 2009). These issues are among the main challenging tasks that face medical information retrieval systems (Goeuriot et al., 2016, Palotti et al., 2016). Another important factor that affects the quality of the returned results by MIR systems is embodied in the richness and domain coverage of the exploited medical semantic resources; which provide formal and explicit specifications of shared medical conceptualizations (Díaz-Galiano et al., 2007, Koopman et al., 2016). Several semantic resources and classification systems have been developed in the medical domain. Examples of such resources are: the Unified Medical Language System (UMLS)(Bodenreider, 2004), Medical Subject Headings (MeSH)

(Lipscomb, 2000) terms, Systematized Nomenclature of Medicine - Clinical Terms SNOMED-CT (Lee et al., 2013), Logical Observation Identifier Names and Codes LOINC (Vreeman et al., 2010), DRUG, Gene and Human Disease ontologies (Consortium, 2011, Hanna et al., 2013, Kibbe et al., 2014), ICD-10 standard (Coustasse and Paul III, 2013), and Pubmed (Malhotra et al., 2015). For more details on existing medical semantic resources please refer to the bioportal[1] gate.

Though the exploitation of such resources has proved to be more effective than conventional approaches that merely employ the old-fashioned information retrieval techniques, the quality of the result produced by medical resources-based approaches still needs improvement as reported in (Consortium, 2011, Hanna et al., 2013, White et al., 2009). This is mainly because existing medical resources suffer from knowledge incompleteness and semantic heterogeneity issues.

Motivated by these observations, we have proposed to develop an MIR system that wherein we have attempted to enhance the quality of the produced medical-search results by reformulating users' queries based on cooperative decisions obtained from multiple medical semantic resources. Additionally, we attempt to integrate multiple medical semantic resources and evaluate the effectiveness of the proposed system when exploiting an integrated set of heterogeneous resources.

The remainder of this chapter is organized as follows. The background and motivations behind the proposed MedIR system are presented in Section 1.2. In Section 1.3, we have reviewed the literature and identify the main strengths and weaknesses of a number of

[1] https://bioportal.bioontology.org/ontologies Retrieved on 02-Feb-2018

MIR systems that are related to our proposed approach. Section 1.4 provides the research steps. Section 1.5 defines our contributions and clarifies the obstacles that we attempt to overcome through the proposed MedIR system. Finally, the structure of our thesis is introduced in section 1.6.

1.2 Motivations

In the medical domain, the search process is more complicated than other fields of information retrieval since users need to formulate their search queries using the right medical terms. For instance, for laypeople, they are usually unaware about the correct medical terms to precisely describe a type of diseases or diagnosis; as a result, the retrieved results may not provide them with the desired information (Stanton et al., 2014). Many attempts have been made to address this issue through the exploitation of medical resources and knowledge bases. In this context, these resources are used to analyze queries and organize the returned results by MIR systems. However, we can find plenty of these resources and it is not clear which among the used resources produces the best results, and what are the contexts in which one resource should be preferred against others.

In light of these problems, we propose building a MIR system that (before carrying out the search process) reformulates users' queries to reduce the gap between how laypeople formulate their needs and what they actually mean by the formulated query, and also to precisely address their information needs in the medical domain. Moreover, the proposed system will take advantage of the available medical resources and knowledge bases to improve the quality of the retrieved results.

1.3 Problem Statement and Research Questions

As we have discussed in the previous section, information retrieval in the medical domain has several problems including the semantic gap and mismatch among users' queries and medical documents terminology due to the frequent use of medical acronyms, abbreviations, and synonyms. The semantic gap illustrates the difference among different descriptions of an object in different representations (Tibi et al., 2017). One of the main objectives of MIR is to fill the semantic gap between the queries and documents and to go beyond keywords matching. Moreover, current MIR 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, traditional IR methods such as keywords matching fail to retrieve semantically similar documents (Hliaoutakis et al., 2006). It is fairly easy for humans to understand exactly semantically similar terms or semantically related terms, but difficulties arise when we need MIR systems to determine semantically similar and closely related terms. Determining semantically similar and semantically related terms can improve the effectiveness of MIR systems by determining closely related diseases, treatments and symptoms. These problems have raised new challenges in the field of MIR, especially given that medical documents and electronic medical records are expanding.

Even though there are many researches have proposed for improving the effectiveness of medical information retrieval systems, they still need more improvements as presented earlier.

In this section, we present the research questions that we have endeavored to examine and address during our research work.

- What are the strengths and weaknesses of the methods and techniques that are employed by current MIR systems?

In order to answer this question, we have conducted a comprehensive analysis of current MIR systems and proposed approaches and studied the features of the implemented methods /techniques for each system.

- How to reduce the gap between the laypeople queries and the related documents in the medical domain?

To tackle this issue, we have proposed a query reformulation technique using multiple integrated medical semantic resources for enriching the users' queries with semantically related medical concepts that properly describe the user need and fill the gap between the query terms and the medical documents terms. Additionally, we have proposed a query terms classification technique using medical semantic resources to classify query terms into medical and supportive terms and assigning a higher weight for medical terms.

1.4 Research Steps

The following points introduce the main steps that we have carried out during our research work:

- **Processing Users' Queries:**

At this phase, we have used conventional natural language processing (NLP) techniques to process user's queries. We start by removing all punctuation marks from the user’s query. Then, we remove stop words based on a pre-defined list such as: a, the, an,…etc. After that, the query terms are stemmed using Porter stemmer (Willett, 2006). Finally, we use the NLP n-grams tokenization technique to tokenize the input text into n-gram tokens of lengths from 1 to 3. More details on this phase are provided in section 4.1.1.

- **Medical Acronyms and Medical Terms Recognition**

During this phase, an automatic extraction of query components such as: medical acronyms, medical abbreviations, medical terms, and other supportive terms is carried out. An automatic expansion of the synonyms of medical query terms is also carried out during this phase. To extract medical acronyms, abbreviations, and synonyms we use the Unified Medical Language System lexicon (UMLS lexicon) (Huang et al., 2005) as an external medical semantic resource. The query is then automatically reformulated by incorporating all of the full representations of the extracted acronyms and abbreviations, and also by adding the extracted synonyms to the reformulated query. We also used UMLS Metathesaurus medical semantic resource for classifying query terms into medical terms and supportive terms.

- **Matching Queries to their Corresponding Medical Documents and Ranking Related Documents:**

In this phase, we have discussed the proposed matching and ranking formulas used to improve the effectiveness of the proposed MedIR system. The key idea of our proposed ranking technique is to assign higher weights for medical query terms over other supportive terms, and also over all terms added to the original user query from external medical semantic resources during the previous phase. We use the vector space model (VSM) that is usually known as cosine similarity model (Larson, 2010) for finding the similarity between the user's query and the medical documents in the document collection. We will discuss in details our proposed terms reweighting method in next sections.

- **System Evaluation**

To validate the efficiency and effectiveness of the proposed MIR system, we have used well known precision quality indicator (specifically, we use P@10 indicator as discussed in section 5.2). Moreover, we have compared the produced results by the proposed system with the best three CLEF participant teams who used the same dataset that we used in our experiments for evaluating their proposed systems (more details about those (more details about those state-of-the-art systems provided in section 5.3). We used CLEF e-Health2014 dataset that comprises of around one million medical documents and fifty-five testing queries (Goeuriot et al., 2014) for system evaluation.

1.5 Contributions

We summarize the main contributions of our research work as follows:

1. Employing a Semantically-enhanced Inverted Index for indexing medical documents. Where the employed index captures the hidden semantic dimensions that are encoded in the text of each medical document. An inverted index consists of a dictionary of terms called vocabulary or lexicon along with a list that indicates which documents the term occurs in. Some statistics such as the term frequency in each document and the document frequency for each term also can be added to the inverted index (Larson, 2010)

2. Reformulating the given users’ queries through enriching them with semantically-related medical concepts using external medical semantic resources to minimize the gap between laypeople queries and medical documents terminology.
3. Classifying user query terms into medical and supportive terms by exploiting medical semantic resources and assigning a higher weight for medical terms against other query terms.

1.6 Structure of the Thesis

The rest of this thesis is organized as follows. In Chapter 2, we present a comprehensive comparative analysis of existing MIR systems. A general overview of the architecture of MedIR system is presented in chapter 3. We introduce a detailed description of the techniques and methods that we have utilized in the MedIR system in Chapter 4. Chapter 5 presents the conducted experiments and discussion in addition to the evaluation of the effectiveness of the proposed system. In this chapter, we also have compared the result produced by our system and three of the recently proposed MIR systems. Finally, in Chapter 6, we have discussed the conclusions and outline the future extensions of our research work.

Chapter 2 - Literature Review

2.1 Introduction

For this reason, thousands of people have adopted the practice of searching the web for information related to their health and the health of their family and friends. Due to the vast amount of information on the Web, this practice is not always efficient. One of the reasons is the difference in the knowledge that each user has about health topics. Several approaches have been proposed for filling the semantic gap between users' queries and medical documents to improve the effectiveness of MIR systems and increase users' satisfaction. By reviewing existing MIR systems, we can clearly find that they have employed different techniques and approaches to cope with the problem. In this Chapter, we also discuss the existing techniques, and provide a comprehensive analysis of the work that has been done by other researchers and developers in the field of MIR and Query Reformulation. Before going through discussing state-of-the-art MIR systems and introducing our proposed approach in details, we provide some background about the main concepts in the information retrieval that will help for better understanding for the rest of the thesis and related work.

2.2 Theoretical Background

Before proceeding to present the details of the methods and techniques used in the proposed system and state-of-the-art systems, we formally define – in the context of our work – the terms “Information Retrieval’’, "Semantic Information Retrieval", “Semantic Resource”, and “Query Reformulation”.

- **Definition 1: Information Retrieval (IR)**

The term Information Retrieval (IR) was adopted by Christopher D. Manning who defined it as:

".... Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (Mogotsi, 2010).

- **Definition 2: Semantic Information Retrieval**

One of the major drawbacks of IR systems is the use of bag-of-words (keywords) information retrieval models that could return imprecise and insufficient results; due to their dependencies on specific keywords existed in the search sentence or query. The semantic IR aims to fill the gap between query-document pairs and overcome the keywords problem by using the conceptual representation for both queries and documents by exploiting existing semantic resources rather than simple keywords representation

- **Definition 3: Semantic Resource**

A semantic resource SR is quadruple, SR = {C, P, I, V} where:

C represents the set of concepts that are defined in the SR. The hierarchical relationship between concepts of the set C is a pair (C, ≤), where the symbol ≤ is an order relation on C x C. We call ≤ the sub-concept relation. P represents the set of properties defined over C. I is the set of individuals also called instances of the concepts in the SR. V is the set of values defined over P.

- **Definition 4: Query Reformulation**

Query reformulation is the process of enriching the original user query with additional terms or deleting some existed terms in order to improve the information retrieval performance (Rieh, 2006). The main objective of the query reformulation techniques is to improve search results for laypeople by reducing the gap between the words used by the user to express his/her need in the initial query and the words or concepts used in the related documents.

2.3 Approaches/ Techniques Utilized by MIR Systems

Many Query Reformulation (QR) approaches and techniques have been proposed for addressing the MIR challenges (Goeuriot et al., 2016) and improving the effectiveness of the medical search results and increase laypeople satisfaction who is searching in this context. In this section, we classify these techniques into two main approaches: query reformulation using external semantic resources (Medical/Non-medical) (Choi and Choi, 2014, Shen et al., 2014, Stanton et al., 2014, Zghal and Moreno, 2014) and query reformulation using pseudo relevance feedback (PRF) technique based on local collection (Oh and Jung, 2014, Yoo and Choi, 2010, Thesprasith and Jaruskulchai, 2014, Lu, 2015). In the remaining parts of this section, we discuss the major drawbacks and limitations that are associated with each approach.

A. Query Reformulation Using External Semantic Resources

Authors in (Zghal and Moreno, 2014) proposed a query reformulation technique to improve the effectiveness of the semantic MIR systems. They proposed a query reformulation approach where the query reformulated via two steps: In the first step, the users' query submitted to MeSH[2] medical semantic resource and reformulated by adding the related general extracted super-classes to the original query. While in the second step, the previously reformulated query submitted to Yahoo Boss crawler[3] to query the Web and gather its related documents and web pages. The key sentences in the retrieved results that contain the original users' query terms extracted from the retrieved results and submitted to the Stanford Part of Speech (POS) tool[4] to identify the concepts exist in them and enriching the reformulated query resulted from the first step with those concepts. The final reformulated query that enriched with semantically related concepts, submitted to PubMed[5] -a well-known MIR system- to evaluate the proposed approach by comparing results obtained by the original user query and by the reformulated query. The proposed system was evaluated by a selected group of Biology and Medicine students, they submitted their own queries and their related reformulated queries resulted from the proposed approach to PubMed and compared the effectiveness of the retrieved results in both cases. Although the results of the evaluation phase were satisfactory as authors conclude, the proposed approach need to be tested using real medical dataset (such as CLEF, TREC,..etc.) that provide queries and relevance judgments created by medical professionals for evaluation purpose to prove its effectiveness. In (Choi and Choi, 2014) the authors proposed a concept-based query reformulation model using selective query concepts. They used CLEF eHealth2014 dataset in their experiments where each query in the dataset has a related 'discharge summary report' as an example of what the relevant results of each query should be like. They used UMLS medical semantic resource to

[2] https://www.ncbi.nlm.nih.gov/mesh Retrieved on 26-MAR-2018
[3] https://developer.yahoo.com/search/boss/ 26-MAR-2018
[4] http://nlp.stanford.edu/software/tagger.shtml 26-MAR-2018
[5] https://www.ncbi.nlm.nih.gov/pubmed/ 26-MAR-2018

extract and expand medical concepts in the user query, and then they ignore all extracted concepts that didn't exist in the query related discharge summary. Their proposed system achieved slight improvement on the effectiveness of MIR systems for two reasons: 1) The indexing method they used for document collection was very simple without using the advanced NLP techniques such as compound terms and stop words removal during the indexing process. 2) They restricted the expansion scope with the content of the query related discharge summary report provided with the dataset, while this report provided as an example of the related results only. The authors in (Shen et al., 2014) proposed a "bag of concepts" retrieval model where they extracted the medical concepts from the users' query using UMLS metathesaurus and reformulated the original query by removing all query terms not identified by UMLS, and adding related extracted terms to the selected concepts. The proposed system was evaluated by developing a prototype and conducting experiments using CLEF eHealth 2014 dataset. The experimental results were far from what authors expected, only a slight improvement achieved over the traditional bag of words retrieval model. The low effectiveness of their approach -according to our opinion- may refer to the following factors: 1) the ignorance of all non-medical terms in the original user query in their retrieval model, and 2) the loss of some medical concepts exist in the original query since they didn't identify by UMLS during concepts extraction. In (Stanton et al., 2014) the authors studied the contexts in which a non-expert uses many words to describe a symptom instead of using the appropriate medical terms. The authors proposed a supervised approach to link searched queries to medical concepts that can be mapped to a specific disease. They used the professional definition of diseases from medical semantic resources to map the user query into a professional query that consists of medical concepts. Their approach achieved an improvement in mapping symptoms to the proper relevant disease, but they didn't consider other query types where the user may require information about any medical data (not symptoms and diseases) such as laboratory test, medical devices, etc. A semantic query reformulation technique was introduced in (Kim et al., 2015) by detecting and mapping terms in users' query to concepts in UMLS and incorporating context data based on two approaches. The first approach depends on PubMed query logs by extracting terms that frequently appear near query conceptual terms to clarify medical concepts. While the other approach depends on the predefined user profile, which contains set of terms that represent user interest. Terms in the user profile compared with input query terms. If there were a good match between query terms and profile terms, a higher weight assigned to the matched query terms. After query reformulation, term weight assigned to each term in the query to affect documents ranking by giving higher weight to related terms from users' profile, then conceptual terms and the lower for context terms from PubMed log. Two evaluation methods were used to measure the effectiveness of the proposed system, objective evaluation using OHSUMED dataset (Freitas ‐ Junior et al., 2006) to evaluate concept-based query reformulation and subjective evaluation by involving human assessment of the results considering user profile. An improvement with the precision and recall achieved with different values vary across queries. We expect that the main reason for the low improvement of their proposed approach is that they didn't consider the relationships (synonyms, hyponyms, acronyms, and abbreviation) between original query terms and their related terms in the reformulated query.

B. Query Reformulation Using Pseudo Relevance Feedback (PRF) technique based on local a collection

Pseudo Relevance Feedback (PRF) is a technique for query reformulation consists of modifying the initial query of the user by adding some terms obtained from the highest-ranked documents retrieved in the IR system by the initial user query (Abderrahim, 2014). Many researches have been conducted in the context of MIR using PRF technique. Examples of these research are (Oh and Jung, 2014, Yoo and Choi, 2010). For instance, the system proposed in (Oh and Jung, 2014) presents a multiple-stage method which focused on utilizing various retrieval techniques rather than exploiting utilizing external resources and NLP techniques. In the first stage, they reformulated the users' query by adding the full representations of abbreviations existed in the user query. To do that, They extracted pairs of abbreviation and corresponding full representation with an occurrence count using simple rule-based extraction method (Schwartz and Hearst, 2002) from the entire collection. Then, a query model is estimated by incorporating words from the full representations of an associated abbreviation. In the second stage, authors reformulated the query using the discharge summary related to each query that provided with the used dataset - CLEF eHealth 2014 (Goeuriot et al., 2014) - and may contain hidden information related to the query but not mentioned in the query itself. In the final stage, the query is reformulated using PRF where the top-ranked documents retrieved from the previous stage considered being relevant and useful for reformulating the user query. Their proposed approach showed slight improvement on the system effectiveness resulted from using abbreviations and discharge summaries, while the use of PRF didn't have a serious impact on results effectiveness as we noticed by observing the results retrieved from each run of their experiments. Authors in (Yoo and Choi, 2010) proposed a query reformulation approach that based on PRF model, The key idea of their approach was to test the effect of using PRF model with different terms ranking methods on improving the effectiveness of MIR systems. They tested the effect of using the following six term ranking methods on the system performance: the Rocchio weight (Carpineto et al., 2001) based on the Vector Space Model, Kullback–Leibler Divergence (KLD) (Carpineto et al., 2001) based on information theory, Robertson Selection Value (RSV) (Carpineto et al., 2001), CHI-squared (CHI2) (Carpineto et al., 2001), Expected Mutual Information Measure (EMIM) (Losee, 2001) based on probabilistic distribution analysis, and Local Context Analysis (LCA) (Xu and Croft, 2000) utilizing co-occurrence with all query terms. Their experiments showed that using LCA terms ranking technique had the best effect on results effectiveness. They concluded that their approach made a slight improvement and that the MIR stills a challenging issue that needs more and more improvements. In (Lu, 2015) the author proposed a query reformulation approach using PRF model with two terms ranking models: Kullback-Liebler Divergence(KLD) model and parameter-free Bose-Einstein statistics-based (Bo1) model. The purpose of this research was to reduce the gap between users' query and document collection in medical the domain by enriching the user query with additional terms selected from the top-k ranked documents retrieved from the initial retrieval process. In their experiments, they tried to find the best documents count and the best expansion terms count that achieves significant improvement on MIR systems effectiveness. After experiments, they concluded that the use of Bo1 model with 3 feedback documents and 10 expansion terms are better than the initial retrieval process but are not statistically significant, and they find that the performance of Bo1 model is better than KLD model. Authors of (Thesprasith and Jaruskulchai, 2014) proposed query reformulation approach using non-traditional PRF paradigm. They reformulate the original query using results of external PRF instead of local PRF. Their approach consists of the following steps:

- **Indexing Process and Initial Retrieval Model**

  They used Lucene[6] retrieval tool for documents indexing and for initial retrieval. The built-in vector space model (known as cosine similarity model) exists in Lucene used as their initial retrieval model.

- **Expansion Source Selection**

  To select the external expansion resource to be used as PRF resource, authors perform an initial retrieval in three TREC standard sub-collections such as med (Voorhees and Harman, 2000), OHSUMED(Hersh et al., 1994), and Genomics 2004 (Hersh et al., 2006) using 5 training queries and they select the collection that retrieved more documents to be the used expansion resource in next steps. Results of the five training queries from the Genomic collection were more than OHSUMED and med where they believe that more documents retrieved, provide more useful expansion terms. For this reason, the Genomics 2004 document collection used as external expansion resource for PRF model.

- **Expansion Term Selection**

  They select terms co-occur more often in the selected expansion resource - Genomics 2004 - set for expansion. Their term selection procedure can be described as follow: First, They retrieved initial documents using Lucene tool

[6] http://lucene.apache.org/core/ Retrieved on 07-May-2018

from the Genomics collection. Second, top-k documents that contain any query terms were included in the PRF-expansion set. Third, terms in title and abstract part of the PRF-expansion set were selected based on term frequency as candidate terms set. They did some experiments to find the best number of documents and terms to be used in their approach, and they found that the top-8 co-occurred terms in the top 19 expansion documents is the best selection.

- **Re-weighting Method**

  They used Rocchio's formula (Abdou and Savoy, 2008) for terms re-weighting. However, they modify the terms weighting to assign a higher weight for original query terms.

The proposed approach showed an improvement from their baseline but the improvement was insignificant as authors concluded. From our opinion, this may refer to the selected expansion resource (Genomics 2004) since this document collection contains genomics information not medical or clinical information.

The following table summarizes the available MIR systems that we have introduced in the previous sections.

Table 1: Summarization of the studied MIR systems

	System/ Approach	**Category**	**Implementation techniques/approaches**
1	A system for information retrieval in a medical digital library based on modular ontologies and query reformulation	Query Reformulation Using External Semantic Resources	Query reformulation via Two stages using multiple resources (medical and non-medical) : - MeSH (medical) - Yahoo Boss crawler (non-medical) - Stanford Part of Speech (POS) tool (non-medical)
2	Exploring Effective Information Retrieval Technique for the Medical Web Documents	Query Reformulation Using External Semantic Resources	Concept-based query reformulation model based on a selective query concepts using UMLS metathesaurus medical

			semantic resource - CLEF eHealth2014 dataset used for system evaluation
3	An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@ CLEF2014 eHealthTask 3	Query Reformulation Using External Semantic Resources	A "bag of concepts" retrieval model that used the UMLS metathesaurus medical semantic resource. - CLEF eHealth2014 dataset used for system evaluation
4	Circumlocution in diagnostic medical queries	Query Reformulation Using External Semantic Resources	Search queries linked to medical concepts that can be mapped to a specific disease using medical semantic resource.
5	Exploring Context-Sensitive Query Reformulation in a Biomedical Digital Library	Query Reformulation Using External Semantic Resources	Search queries reformulated using UMLS metathesaurus and PubMed query logs. - A weight assigned to query terms based on user profile.
6	Share/clef eHealth evaluation lab 2014, task 3: User-centered health information retrieval	Query Reformulation Using Pseudo Relevance Feedback (PRF) technique based on local collection	Multiple-stage query reformulation approach: 1. Adding full representation of abbreviations based on simple rule-based method. 2. Search query reformulated using discharge summary in the used dataset. 3. Search query reformulated using PRF where the top-ranked documents retrieved from the previous stage.
7	On the query reformulation technique for effective MEDLINE document retrieval	Query Reformulation Using Pseudo Relevance Feedback (PRF) technique based on a local collection	Search query reformulated using PRF model from local collection. - six terms ranking methods tested to select the best for query terms ranking
8	Employing Query Expansion Models to help Patients Diagnose Themselves	Using Pseudo Relevance Feedback (PRF) technique based on a local collection	Search query reformulated using PRF by focusing on finding the best documents count and terms count used in the expansion process
9	CSKU GPRF-QE for Medical Topic Web Retrieval	Using Pseudo Relevance Feedback (PRF) technique based on a local collection	Search query reformulated using external non-traditional PRF based on three expansion Sources: med, OHSUMED, and Genomics.

2.4 Summary

The aim of this chapter was to introduce a literature review of the existing MIR systems.

We have explained the approaches/ techniques utilized by MIR systems including

Reformulating Users' Queries Using External Semantic Resources (either they are medical or non-medical semantic resources) and Reformulating Users' Queries Using PRF technique based on the local collection. From our review of the conducted research in the context of medical information retrieval, we infer that the use of external semantic resources specifically medical ones is achieving better improvements than other proposed approaches. So, we decide to use an integrated external medical semantic resources for reformulating the user's queries in order to improve the effectiveness of the existing MIR systems.

Chapter 3 - Query reformulation in the context of Medical Information Retrieval

3.1 Introduction

In this chapter, we introduce our proposed medical information retrieval approach. First, we present a general overview of our proposed MedIR system. Then, we clarify the overall architecture of the proposed system. Finally, we summarize this chapter in section 3.3.

3.2 General Overview of the MedIR System

In this section, we present a general overview of MedIR system. In our research work, we propose an approach that exploits query reformulation techniques and a set of integrated medical semantic resources for improving the effectiveness of MIR systems, by enriching laypeople queries with semantically related concepts that best describe their information need. Figure 1 below shows an overview of the MedIR system.

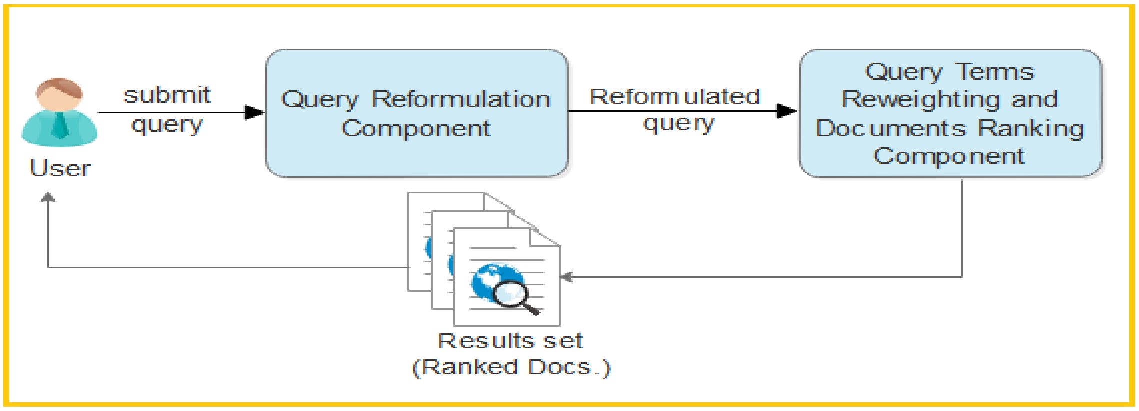

User
submit query
Query Reformulation Component
Reformulated query
Query Terms Reweighting and Documents Ranking Component
Results set (Ranked Docs.)

Figure 1: General architecture of MedIR system

As depicted in figure1, the MedIR system comprises two major components (Query Reformulation component and Query Terms Reweighting and Documents Ranking component). In the next sections, we will present the details of each component.

3.2.1 Query Reformulation Component

We have constructed the first component of the MedIR system based on employing a set of NLP techniques for query preprocessing and an integrated medical semantic resources for efficiently reformulate the user's query in a way that best describes the user needs. Figure 2 describes the overall architecture of the Query Reformulation Component.

As shown in Figure 2, the first component of our system comprises several modules that are organized as follows:

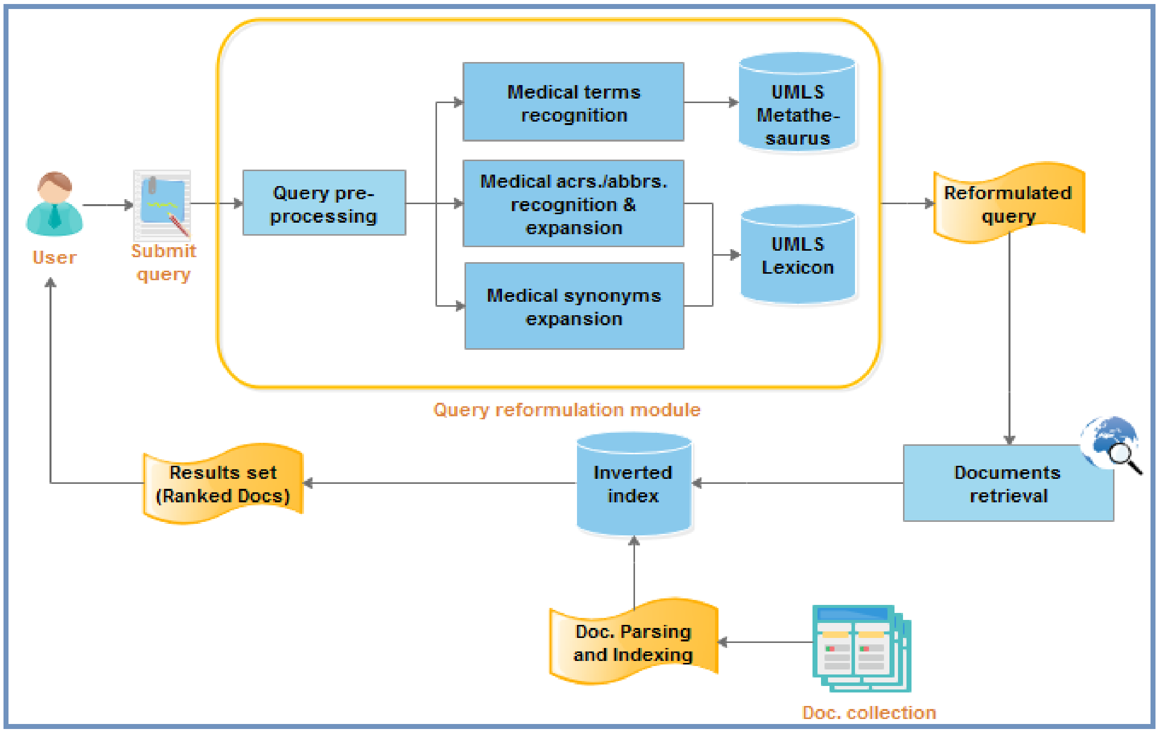

Medical terms recognition
UMLS Metathe-saurus
User
Submit query
Query pre-processing
Medical acrs./abbrs. recognition & expansion
Reformulated query
UMLS Lexicon
Medical synonyms expansion
Query reformulation module
Results set (Ranked Docs)
Inverted index
Documents retrieval
Doc. Parsing and Indexing
Doc. collection

Figure 2: General architecture of the proposed query reformulation component

First, the original user query is pre-processed using ***Query Pre-processing Module*** based on performing the following NLP techniques (more details are provided in section 4.1.1):

- Case folding
- Punctuation marks removal.
- Stop words removal

- N-gram text tokenization
- Tokens stemming

By performing the above-mentioned steps, we have converted the original user query into a normalized query that contains only semantic significance terms with minimal lexical variants. Next, a Medical ***Acronyms / Abbreviations Recognition & Expansion module*** is used to extract the medical acronyms and abbreviations existed in the processed user's query resulted from the previous module and expand them with their full representations. After that, the ***Medical Synonyms Expansion module*** defines the concepts in the processed user's query that have some related medical synonyms in a selected medical semantic resource and expand them with their related synonyms. The last module of the MedIR system, namely the ***Medical Terms Recognition module*** extracts the medical concepts (either they are unigrams, bigrams or trigrams) that exists in the user's query using another external medical semantic resource and adding them to medical terms classification. The output of this module is a reformulated query enriched with the related medical semantic concepts that best describe user needs, and a classification of the original query terms into the following query components (medical acronyms, medical abbreviations, medical concepts, and supportive terms).

3.2.2 Query Terms Reweighting and Documents Ranking Component

In this component, we utilize both the reformulated query and the classifications of the query terms - resulted from the previous component - to improve the effectiveness of medical search results using the proposed terms reweighting technique and documents ranking approach.

Figure 3 depicts the overall architecture of the MedIR system after integrating the new components for the query terms reweighting and documents ranking.

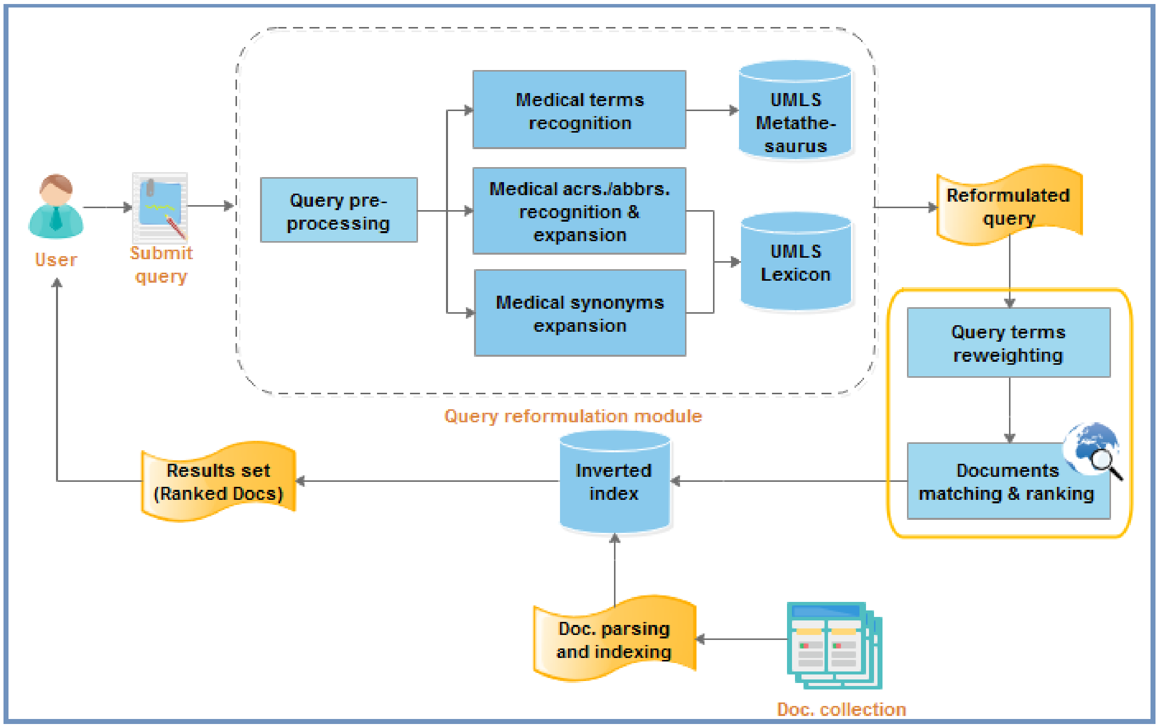

Medical terms recognition
UMLS Metathe-saurus
Query pre-processing
Medical acrs./abbrs. recognition & expansion
UMLS Lexicon
Medical synonyms expansion
User
Submit query
Reformulated query
Query terms reweighting
Query reformulation module
Results set (Ranked Docs)
Inverted index
Documents matching & ranking
Doc. parsing and indexing
Doc. collection

Figure 3: General architecture of the proposed query terms reweighting and documents ranking component

The Key idea of this proposed component is to give more importance for medical query terms over other general terms in the user's query. This has been done by assigning a higher weight to the following query components (medical acronyms, medical abbreviations, medical concepts and the added medical synonyms) over other remaining components in the reformulated query (supportive terms and full representations of both acronyms and abbreviations). To find the related documents to the users' query, We used keywords matching by finding all documents that contains any term in the reformulated query. For the related documents ranking process, we use the vector space model (Larson, 2010) that employs the term frequency-inverse document frequency (tf-idf) (Christopher et al., 2008) weighting scheme for assigning weights to query terms. We enhance the well-known information retrieval models used in order to cope with our goal in improving the effectiveness or MIR systems. In section 4.2, We will discuss the modified formulas proposed in our approach in more details.

Experimental instantiation of the proposed system has been installed to demonstrate the effectiveness of this component and other proposed components in our approach. We provide more details on the conducted experiments in chapter 5.

3.3 Summary

Our goal in this chapter was to provide a general overview of our proposed MIR system and to clarify the overall architecture of the proposed system. In addition, we elaborated that our proposed system has two main components (Query Reformulation and Query Terms Reweighting and Documents Ranking). In addition, we clarified the role of each component in improving the effectiveness of existing MIR systems.

Chapter 4 - MedIR System Implementation

In this chapter, we present the implementation details of the MedIR system. First, we present the details that pertain to the first version of the system's prototype in section 4.1. Then, in section 4.2, we describe the updated version of the system's prototype and detail the features of the new components. Finally, we summarize this chapter in section 4.3.

4.1 Development Details of the Query Reformulation Component

During this phase, we have implemented the first prototype of the MedIR system through incorporating a set of NLP techniques and an integrated medical semantic resources that combine two main knowledge bases: UMLS lexicon (Huang et al., 2005) and UMLS metathesaurus (Bodenreider, 2004) (see section 4.1.3). The NLP techniques and the semantic resources are exploited to enrich the original user query with semantically related concepts that best describes the user needs and to classify query terms into the following components:

- Acronyms (Keith and Burridge, 1991): terms that are formed from the initial letters of some longer names and is pronounced differently than the full representation (such as: ARV that stands for 'Adelaide River Virus', 'Average Rectified Value' or other medical terms).
- Abbreviations (Keith and Burridge, 1991): are terms that are written differently from their full representations, but are pronounced the same (such as: Abd that stands for 'Abduction').
- Medical Terms: are terms that can be mapped to medical concepts in medical semantic resources (such as: aortic).

- Other Supportive Terms: any other terms0 in the user query that could not be classified as acronyms, abbreviations, or medical terms (such as: replacement, status, etc.).

The prototype system comprised several modules as detailed below:

4.1.1 Query Pre-processing Module

When a user submits a medical query, this module is employed to automatically process the submitted query using a set of NLP techniques in order to be used in the following modules to achieve a better improvement regarding query terms classification. The used NLP techniques including:

- Case folding by reducing all letters into lower case.
- Punctuation marks removal.
- Stop words removal: we pre-defined a list of stop words that contains 566 stop word such as (a, the, we, in, as,...etc.) that have no semantic significance and do not contribute in the actual meaning of the users queries. Such stops words are automatically removed to enhance the system's performance and effectiveness during query reformulation and matching process.
- N-gram text tokenization: we carry out this step in order to tokenize the query into unigrams, bigrams and trigrams tokens.
- Tokens stemming using Porter stemmer (Willett, 2006).

The following example clarifies the process of query pre-processing:

Example 1: Step 1 - Query Pre-processing

- User Query (Uq)**:**

$Uq1$: MRSA and wound infection, and its danger (QTRAIN2014.1 of CLEF e-health2014 dataset (Goeuriot et al., 2014)).
$Uq2$: Peptic Ulcer disease (qtest2014.35 of CLEF e-health2014 dataset (Goeuriot et al., 2014)).

In this example, we convert $Uq1$ and $Uq2$ into lists of processed n-grams as follows:

$Uq1$ n-grams:

- List of unigrams (U_t1): [mrsa, wound, infect, danger].
- List of bigrams (B_t1): [mrsa wound, wound infect, infect danger].
- List of trigrams (T_t1): [mrsa wound infect, wound infect danger].

$Uq2$ n-grams:

- List of unigrams (U_t1): [mrsa, wound, infect, danger].
- List of bigrams (B_t1): [mrsa wound, wound infect, infect danger].
- List of trigrams (T_t1): [mrsa wound infect, wound infect danger].

Once the original user query converted into unigrams, bigrams, and trigrams tokens, we employ the following three modules:

1. Acronyms / Abbreviations Recognition and Expansion Module: to identify, extract, and expand medical acronyms and abbreviations into their full representation.

2. Medical Terms Recognition Module: to identify and extract the medical terms in the submitted user query.

3. Medical Synonyms Expansion Module: to extract the synonyms of the medical terms identified in the previous module and add them to the expansion terms list.

The following three Sections 4.1.2, 4.1.3 and 4.1.4 will discuss in details the proposed three modules mentioned above respectively.

4.1.2 Acronyms / Abbreviations Recognition & Expansion Module

During this module, an automatic extraction and expansion of query components such as: medical acronyms and medical abbreviations are carried out based on the n-gram tokens resulted from the previous query pre-processing module. To do this, we use the UMLS SPECIALIST lexicon (known as UMLS lexicon) (Huang et al., 2005) provided by the National Library of Medicine (NLM[7]). UMLS lexicon is one of the richest available sources of medical lexical information and the most complete lexicon than any others for purposes of analyzing medical text (Huang et al., 2005, Szolovits, 2003). In our proposed approach, we use the ACRONYM relational to do the following:

1. Extract medical acronyms and abbreviations existed in the user query.
2. Expand the extracted acronyms and abbreviations into their full representations and use them to reformulate the original user query.

The ACRONYM table contains 175,080 different medical acronyms and 86,309 different medical abbreviations. Each acronym or abbreviation may relate to many different medical concepts as shown in Table 2:

Table 2 : Example on acronym/abbreviation with their full representations in UMLS lexicon

Acronym/ Abbreviation	Type	Full representations
arv	acronym	average rectified value
arv	acronym	aids associated retrovirus
arv	acronym	avian reovirus
arv	acronym	antiretroviral therapy
arv	acronym	antiretroviral
arv	acronym	adelaide river virus
arv	acronym	antiretroviral drug

[7] https://www.nlm.nih.gov/ Retrieved on 05-Jan-2018

y	abbreviation	young
y	abbreviation	yield coefficient
y	abbreviation	yield

To explain how this module works and what is the output that we obtain after employing it on the user query; we apply this module to the queries in Example 1. Here we show the results:

Example 1: Step 2 - Medical Acronyms/ Abbreviations Recognition and Expansion

$Uq1$:

- List of medical acronyms ($ACt1$): [mrsa]
- List of medical abbreviations ($ABt1$): [].
- List of Expansion Terms ($Et1$): [methicillin resistant staphylococcus aureus].

$Uq2$:

- List of medical acronyms ($ACt2$): []
- List of medical abbreviations ($ABt2$): [].
- List of Expansion Terms ($Et2$): [].

From example 1- step 2, we find that $Uq1$ has one acronym 'mrsa' and no abbreviations. We submit this acronym to UMLS lexicon, and only one related concept found 'methicillin resistant staphylococcus aureus', we add this concept to an expansion terms list $Et1$ that we will use to enrich the original user query and generate the reformulated query $Rq1$.

For$Uq2$, no existence of any acronyms or abbreviations found. Therefore, its related expansion terms list still empty.

After employing this module, the reformulated queries related to user's queries in example 1 will be as follow:

Example 1 - Reformulated Query (Rq):

- $Rq1$: mrsa, wound, infect, danger, methicillin resistant staphylococcus aureus
- $Rq2$: peptic, ulcer, diseas

To expand the medical acronyms and abbreviations terms into their full representation form, we used the following algorithm:

Algorithm 1. Expansion of acronyms and abbreviations recognized into their full representation forms

```
Input: Acr_abbr_list [t1, t2, …,tn]
Output: list of the full representations of all acronyms and abbreviations in the user query.
Temp_list ←⟨⟩;
Exp_list ←⟨⟩;
for i←0; i < Acr_abbr_list.length; i++
    Temp_list = GET_EXP_FROM_LEXICON(Acr_abbr_list [i]);
    for j←0; j < Temp_list.length; j++
        ADD(Exp_list, Temp_list[j]);
    end for
end for
Return Exp_list;
```

4.1.3 Medical Terms Recognition Module

By employing this module, we complete the process of query terms classification and query reformulation. The key idea behind this module is to extract the medical terms existed in the user query. To do this, we use the UMLS metathesaurus medical semantic resource (Bodenreider, 2004) that is a large biomedical thesaurus that is organized by concepts and links together the similar names for the same concept from different semantic resources (nearly 200 resources). We use the following two methods to exploit the UMLS metathesaurus in the process of medical terms extraction:

3. MetaMap tool which maps biomedical texts to the UMLS Metathesaurus. It locates all the UMLS concepts associated with terms in biomedical texts using knowledge intensive method based on symbolic, natural language processing and computational linguistic techniques (Aronson and Lang, 2010).
4. MRDEF relational table that contains UMLS concepts definitions from multiple medical semantic resources. In our approach, we use the 'MSH' source that related to Medical Subject Headings (MeSH[8]) thesaurus and contains 29,244 different concepts.

We submit the user query to MetaMap in order to extract the medical terms in the user query. All query terms that mapped to UMLS concepts are considered to be medical terms. Since MetaMap mainly deals with biomedical texts, it may fail in extracting some medical terms in the lay user's query (which usually contains simple description to express the professional medical concept). To fill in this gap and guarantee the best extraction of all medical terms in the user query, we exploit the medical concepts definition from MRDEF table in the UMLS metathesaurus. The concepts definition expressed by gathering the related attributes that best describe its meaning. After employing this module on the user query, we obtain a list of terms (uni, bi and trigrams) that we add to medical terms classification described above. Finally, all query terms that not existed in acronyms list, abbreviation list or medical terms list will be considered as supportive terms.

If we employ this module to the user queries in Example 1, we will obtain the following results:

[8] https://www.nlm.nih.gov/mesh/ meshhome.html/ Retrieved on 20-Jan-2018

Example 1: Step 3 - Medical Terms Recognition

$Uq1$:

- List of medical terms ($Mt1$): [wound, infect].
- List of supportive terms ($St1$): [danger].

$Uq2$:

- List of medical terms ($Mt2$): [peptic, ulcer, diseas, peptic ulcer, peptic ulcer diseas].
- List of supportive terms ($St2$): [].

From Example 1: Step - 3, we find that $Uq1$ contains two medical terms and one supportive term. While $Uq2$ has five medical terms (unigrams, bigrams, trigrams) without any supportive term.

After employing this module, the reformulated queries related to user's queries in example 1 will be as follow:

Example 1 - Reformulated Query (Rq):

- $Rq1$: mrsa, wound, infect, danger, methicillin resistant staphylococcus aureus
- $Rq2$: peptic, ulcer, diseas, peptic ulcer, peptic ulcer diseas

4.1.4 Medical Synonyms Expansion Module

During this module, we use the medical terms extracted in the previous module and expand them by finding their related medical synonyms in the medical semantic resource. To do this, we use the UMLS lexicon (Huang et al., 2005, Szolovits, 2003) specifically the LEXSYNONYM relational table that contains 5,198 different medical concepts with their synonyms. We use this extracted synonyms to enrich the user query with more concepts for improving the medical search effectiveness.

The following results will be achieved after applying this module to the user's queries in Example 1:

Example 1: Step 4 - Medical Synonyms Expansion

- $Uq1$:
- List of synonyms ($SYt1$): [vulnerat].

$Uq2$:

- List of synonyms ($SYt2$): [digest, ulcu, mal].

From Example 1: Step - 4, we find that $Uq1$ contains one medical term that have synonym 'wound' and it has only one synonym ' vulnerate' that transformed to ' vulnerat' after applying NLP techniques. While $Uq2$ has three medical terms that have synonyms 'peptic, ulcer, disease' that have the following three synonyms respectively 'digestion, ulcus, mal' that transformed to ' digest, ulcu, mal' after applying NLP techniques.

After employing this module, The reformulated queries related to user's queries in example 1 will be as follow:

Example 1 - Reformulated Query (Rq):

- $Rq1$: mrsa, wound, infect, danger, methicillin resistant staphylococcus aureus, vulnerat
- $Rq2$: peptic, ulcer, diseas, peptic ulcer, peptic ulcer diseas, digest, ulcu, mal

To expand the medical terms into their related synonyms from UMLS lexicon, the following algorithm used:

Algorithm 2. Expansion of medical terms into their related synonyms

```
Input: Med_list [t1, t2, …,tn]
Output: list of synonyms for medical terms in the users' query.
Syn_list ←⟨ ⟩;
for  i←0; i < Med_list.length; i++
    if HAS_SYNONYM(Med_list[i]) then
        ADD(Syn_list,GET_SYN_FROM_LEXICON(Med_list[i]));
    end if
end for
Return Syn_list;
```

4.2 Development Details of the Query Terms Reweighting and Documents Ranking Component

During this phase, we discuss the proposed matching and ranking formulas used to improve the effectiveness of our proposed system. We exploit the information extracted from the user query during previous modules to retrieve the best matching documents and ranking them from the most relevant documents to the least ones. As we discuss in the previous modules, each user query is classified into its main components and then enriched with the best semantically related terms that help in best describe the user needs. During query classifications, the following four types of terms are extracted from the original user query:

- Medical acronyms ACt
- Medical abbreviations ABt
- Medical terms Mt
- supportive terms St

We propose to give the same importance to each of the first three classifications in the user query since all of them is a type of medical concepts. But, for the other remaining classification 'supportive terms' we propose to be less important since it has an effect on the query meaning but not primary concepts that the user need to search about. Based on this, we propose to assign a higher weight for the medically related concepts over supportive concepts. In the same context and as we discuss in previous modules, there are two other types of terms added to the user query as an enrichment term. These types are:

- Medical terms synonyms (Synonym terms) SYt .
- Full representations of both acronyms and abbreviations (expansion terms) Et .

In our approach, we propose to give SYt the same weight as other medical terms types (ABt, ACt, Mt) in the original user query and less weight for Et . The main reason for such proposal is that the synonyms provide the exact same meaning as the original term, while the acronyms may have multiple meanings according to the context.

Term frequency-inverse document frequency *Tf-Idf* (Christopher et al., 2008) weighting scheme to assign a weight w for each term t in a document *d* using the following formula:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

(Christopher et al., 2008) **(1)**

Where (idf_t) is the inverse document frequency of the term (t) and is assigned as follows:

$$idf_t = \log_{10}(\frac{N}{df_t})$$

(Christopher et al., 2008) **(2)**

Where:

N is the number of documents in the document collection.

df_t is the number of documents in the collection that t occurs in.

For $tf_{t,d}$, we propose to use the *Normalized-*$tf_{t,d}$ instead of using the basic $tf_{t,d}$ formula. In *Normalized-*$tf_{t,d}$ the term occurrences usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document), to give a measure of the importance of the term t within the particular document *d*. We use Formula (3) to calculate the *Normalized-*$tf_{t,d}$:

$$Normalized - tf_{t,d} = \begin{cases} tf_{t,d} / |\mathrm{d}| & if \quad tf_{t,d} > 0 \\ 0, & Otherwise \end{cases}$$

(Christopher et al., 2008) **(3)**

Where:

|d| is the length of the document *d*.

$tf_{t,d}$ is the number of occurrences for term t in a document *d*.

In our system, we use formula (3) to find the document terms occurrences $tf_{t,d}$, and propose formula (4) to calculate the query terms occurrences $tf_{t,q}$. The key idea of our proposed $tf_{t,q}$ formula, is to reweight the user query terms by giving a higher weight for medical terms *Mt*, medical acronyms *ACt*, medical abbreviations *ABt*, and medical synonyms *SYt*, over other supportive terms *St* and other semantic related terms *Et* used to enrich the user query.

$$tf_{t,q} = \begin{cases} tf_{t,q} / |Uq| & if \quad t \in [Mt, ACt, ABt, SYt] \\ 0.5 \times (tf_{t,q} / |Uq|) & if \quad t \in [St] \\ 1 / |Rq| & if \quad t \in [Et] \end{cases}$$

(4)

Where:

$|Uq|$ is the length of the original user query after applying NLP steps explained in section 4.1.1.

$|Rq|$ is the length of the reformulated query.

In formula (4), we give the medical synonyms *SYt* (that we used to enrich the original user query in section 4.1.4), the same weight as its related medical term *Mt* in the original user query.

To explain our approach, we will apply query terms reweighting formula on Example 1 queries:

Example 1: Step 5 - Query Terms Reweighting

$Uq1$: mrsa, wound, infect, danger
$Rq1$: mrsa, wound, infect, danger, methicillin resistant staphylococcus aureus, vulnerat

$|Uq1|$ = 4
$|Rq1|$ = 6
$tf_{'mrsa',Uq1}$ = 1/ 4 = 0.25 (ACt)
$tf_{'wound',Uq1}$ = 1/ 4 = 0.25 (Mt)
$tf_{'infect',Uq1}$ = 1/ 4 = 0.25 (Mt)
$tf_{'danger',Uq1}$ = 1/2 * 1/ 4 = 0.125 (St)
$tf_{'methicillin\ resistant\ staphylococcus\ aureus',Rq1}$ = 1/ 6 = 0.167 (Et)
$tf_{'vulnerat',Uq1}$ = 1/ 4 = 0.25 (SYt)

$Uq2$: peptic, ulcer, diseas, peptic ulcer, peptic ulcer diseas
$Rq2$: peptic, ulcer, diseas, peptic ulcer, peptic ulcer diseas, digest, ulcu, mal

$|Uq2|$ = 5

$$|Rq2| = 8$$

$$tf_{'peptic',Uq2} = 1/5 = 0.2 \qquad (Mt)$$

$$tf_{'ulcer',Uq2} = 1/5 = 0.2 \qquad (Mt)$$

$$tf_{'diseas',Uq2} = 1/5 = 0.2 \qquad (Mt)$$

$$tf_{'peptic\ ulcer',Uq2} = 1/5 = 0.2 \qquad (Mt)$$

$$tf_{'peptic\ ulcer\ diseas',Uq2} = 1/5 = 0.2 \qquad (Mt)$$

$$tf_{'digest',Uq2} = 1/5 = 0.2 \qquad (SYt)$$

$$tf_{'ulcu',Uq2} = 1/5 = 0.2 \qquad (SYt)$$

$$tf_{'mal',Uq2} = 1/5 = 0.2 \qquad (SYt)$$

As shown in example1- *Uq1*, the medical terms in the original user query and their synonyms in the expanded query take the highest weight over other classifications. Next. the expansion terms (full representations of medical acronyms) take less weight than medical terms. Finally, the supportive terms take the least weight among all reformulated query terms. While in example1- *Uq2*, all the query terms -in the original user query and in the reformulated query - take the same weight since all of them classified as medical terms and medical synonyms. The following algorithm used for query terms reweighting based on query terms classifications.

Algorithm 3. Reweighting the reformulated query terms

```
Input: Rq_terms_list [t1, t2, …,tn], Uq_terms_list [t1, t2, ..., tn],
       Acr_abbr_list [t1, t2, ...,tn], Syn_list [t1, t2, ...,tn],
       Med_list [t1,t2, ..., tn], Sup_list [t1,t2, ..., tn],
       Exp_list [t1, t2, ..., tn]
Output: hashmap of reformulated query terms with their weights.
weight_hmap ←⟨⟩;
term_count = 0;
term_weight = 0;
for  i←0; i < Rq_terms_list.length; i++
    term_count = GET_TERM_COUNT(Rq_terms_list[i]);
    if Med_list.contains(Rq_terms_list[i]) or
```

```
          Acr_abbr_list.contains(Rq_terms_list[i]) or
          Syn_list.contains(Rq_terms_list[i]) then
            term_weight = term_count / Uq_terms_list.length;
    else
            if Sup_list.contains(Rq_terms_list[i]) then
                term_weight = 0.5 * (term_count / Uq_terms_list.length);
            else
                if Exp_list.contains(Rq_terms_list[i]) then
                    term_weight = 1 / Rq_terms_list.length);
                end if
            end if
    end if
    weight_hmap.PUT(Rq_terms_list[i],term_weight);
end for
Return weight_hmap;
```

After calculating *Tf-Idf* for both query and documents terms, the cosine similarity model used to find the semantic similarity between all candidates documents that related to the reformulated user query. We used the following algorithm to find the related documents to the reformulated query.

Algorithm 4. Matching between the reformulated query and their corresponding medical documents

```
Input: Rq_terms_list [t1, t2, …,tn]
Output: list of related medical documents for a given user query
temp_doc_list ←⟨ ⟩;
related_doc_list ←⟨ ⟩;
for  i←0; i < Rq_terms_list.length; i++
    temp_doc_list ← GET_DOCS_FROM_INDEX (Rq_terms_list[i]);
    for  j←0; j < temp_doc_list.length; j++
        if temp_doc_list[j] Not IN related_doc_list then
            ADD (related_doc_list, temp_doc_list[j]);
        end if
    end for
end for
Return related_doc_list;
```

The cosine similarity model deals with both document *d* and user query *Uq* as vectors. Let $\vec{d}$ be the vector representation of *d*, and $\overrightarrow{Uq}$ is the vector representation of *Uq*. To find the similarity between these two vectors, the following formula that attempts to find a relevance score between a given document *d* and a user query *Uq* based on their dot product is used:

$$sim(d,Uq) = \cos ine(d,Uq) = \frac{\vec{d}.\overrightarrow{Uq}}{\left|\vec{d}\right|\left|\overrightarrow{Uq}\right|}$$

(Larson, 2010) **(5)**

Where:

$\left|\overrightarrow{Uq}\right|$ is the length of the user query *Uq*.

$\left|\vec{d}\right|$ is the length of related document *d*.

Experimental instantiation of the proposed system has been done to demonstrate the effectiveness of the proposed approach in improving MIR systems and maximizing lay user's satisfaction in the context of medical search. For more details on the carried out experiments, please refer to chapter 5.

4.3 Summary

The aim of this chapter was to present the methods and techniques that are used in our proposed system. In addition, we have demonstrated that the proposed system has two components. The first component incorporated four modules (*Query Pre-processing Module, Acronyms / Abbreviations Recognition & Expansion Module, Medical Terms Recognition Module, Medical Synonyms Expansion Module*) that mainly focused on classifying user's queries into their components and enriching them with semantically related concepts using the following integrated medical semantic resources:

- *UMLS lexicon*
- *UMLS metathesaurus*

In the second component of the prototype system, a new module for reweighting user query terms was proposed to assign a higher weight for medical terms over non- medical terms in the reformulated query resulted from the first system component.

Chapter 5 - Experimental Evaluation

This chapter describes the experiments that we have carried out to evaluate the techniques of the proposed modules. We have evaluated the effectiveness of our proposed approach by comparing the results produced by our systems' prototype with:

- Baseline run results that we achieved from our experiments using the initial prototype for MIR system that uses the simple inverted index without using any external semantic resources or terms reweighting techniques. The baseline is a measure of process functionality before any change occurs. In Information Retrieval it is a weighting model that counts as a run and allows comparison with the approach applied to verify if the improvement was accomplished (Ruthven and Kelly, 2011).
- Three of the state-of-the-art MIR systems recently proposed in which the authors used the same dataset we use (CLEF e-Health2014).

We have implemented the prototype of the proposed system using Java with PrimeFaces[9] framework and Java Server Faces (JSF) programming languages and conducted the experiments using a PC with Core i7 CPU (2.5GHz), (8GB) RAM and Windows 7 operating system. We have also used Oracle 11g database to build our inverted index and download needed semantic resources locally (UMLS lexicon and UMLS metathesaurus).

The rest of this chapter is organized as follows. We start by describing the dataset used for evaluating our system (CLEF e-Health2014) in Section 5.1. Section 5.2 describes the documents processing and indexing in the document collection, and also describes the

[9] https://www.primefaces.org/ Retrieved on 09-Sep-2017

conducted experiments of the proposed system. Experimental results and discussion are discussed in section 5.3. Finally, we summarize this chapter in section 5.4.

5.1 Dataset (Test Collection)

In order to evaluate the efficiency and effectiveness of our proposed system, we use a CLEF e-Health 2014 dataset (Goeuriot et al., 2014). This dataset provided by the Cross-Language Evaluation Forum (CLEF) eHealth Lab organizers as an evaluation collection for MIR systems. It has a size of 7.2 GB when compressed and approximately 54 GB when extracted and it organized with the following format:

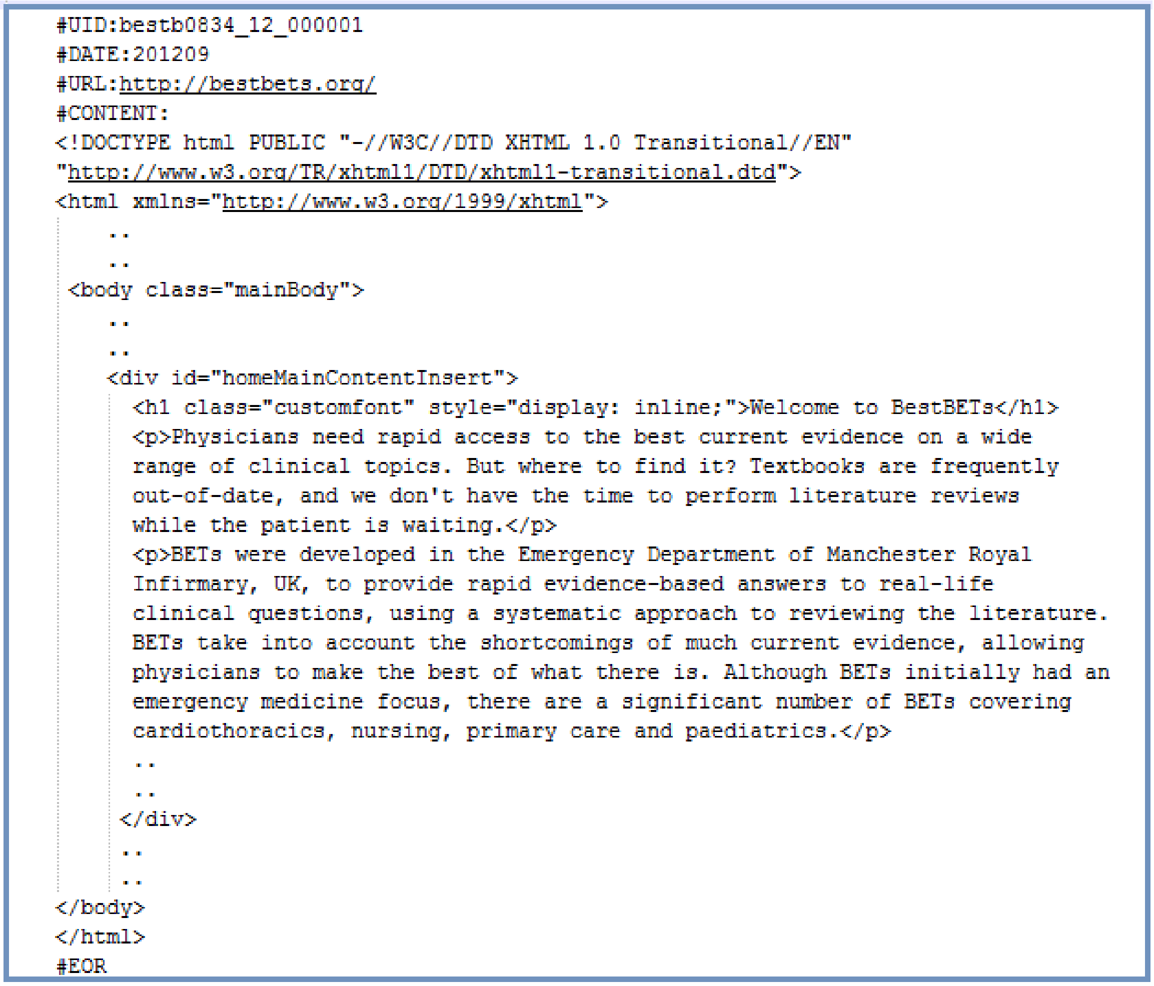
A. Document Collection

The document collection was provided by the Khresmoi project[10] and covers a broad set of medical topics. It consists of automatically crawled web pages from various medical websites, including pages certified by the Health On the Net[11] and other well-known medical websites and databases (Goeuriot et al., 2014). These web pages have a broad range of health topics and are likely to target both laypeople and professionals. It contains around one million documents provided as semi-structured reports in raw HTML format and distributed over 8 .zip files; where each file contains multiple .dat files with different medical topics, and each .dat file contains multiple documents with the following format (see Figure 4):

- #UID: Unique identifier for a document in the collection;
- #DATE: Date the document was obtained;
- #URL: URL for the source of the document;

[10] http://khresmoi.eu/ Retrieved on 02-June-2017
[11] http://www.hon.ch/ Retrieved on 07- June-2017

- #CONTENT: Contains the raw HTML content of web pages.



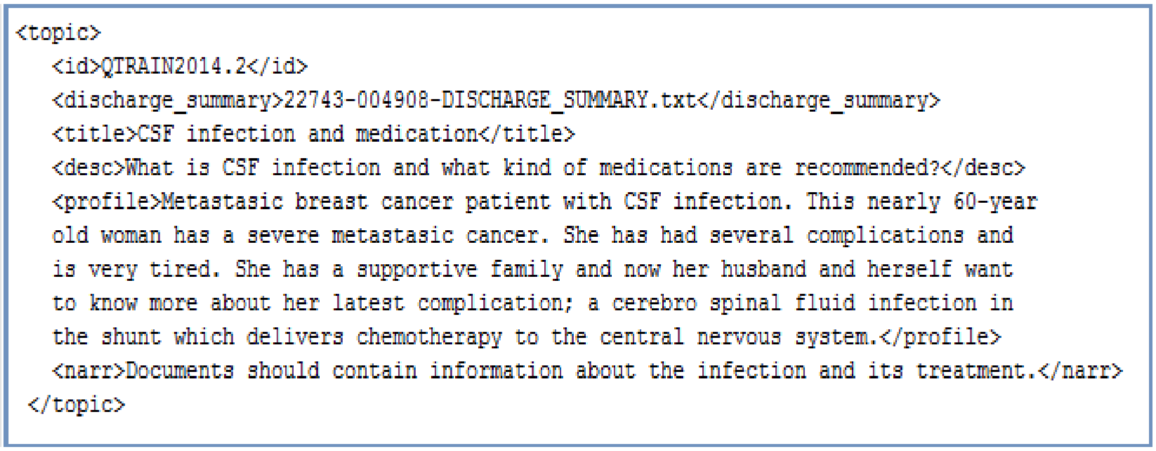
```
#UID:bestb0834_12_000001
#DATE:201209
#URL:http://bestbets.org/
#CONTENT:
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
     ..
     ..
 <body class="mainBody">
     ..
     ..
     <div id="homeMainContentInsert">
       <h1 class="customfont" style="display: inline;">Welcome to BestBETs</h1>
       <p>Physicians need rapid access to the best current evidence on a wide
       range of clinical topics. But where to find it? Textbooks are frequently
       out-of-date, and we don't have the time to perform literature reviews
       while the patient is waiting.</p>
       <p>BETs were developed in the Emergency Department of Manchester Royal
       Infirmary, UK, to provide rapid evidence-based answers to real-life
       clinical questions, using a systematic approach to reviewing the literature.
       BETs take into account the shortcomings of much current evidence, allowing
       physicians to make the best of what there is. Although BETs initially had an
       emergency medicine focus, there are a significant number of BETs covering
       cardiothoracics, nursing, primary care and paediatrics.</p>
       ..
       ..
      </div>
      ..
      ..
</body>
</html>
#EOR
```

Figure 4: Example of a Document from CLEF eHealth 2014

B. Queries

The queries in the dataset are five training queries and fifty test queries that have been created by experts involved in the CLEF e-Health consortium (registered nurses and clinical documentation researchers) and aims to model those used by laypeople (i.e., patients, their relatives or others). The fifty-five queries created based on the main disorders diagnosed in a set of selected patients discharge summaries. Queries are provided in a standard format consisting of the following (Goeuriot et al., 2014) (see Figure 5):

- Title: a short query that expresses the lay people needs.
- Description: more details about the user query in the title.
- Narrative: the expected content of the relevant documents.
- Profile: a brief description of the patient that the query created based on his/her discharge summary.
- Related discharge summary: sample of the information that should be included in relevant documents.



```
<topic>
   <id>QTRAIN2014.2</id>
   <discharge_summary>22743-004908-DISCHARGE_SUMMARY.txt</discharge_summary>
   <title>CSF infection and medication</title>
   <desc>What is CSF infection and what kind of medications are recommended?</desc>
   <profile>Metastasic breast cancer patient with CSF infection. This nearly 60-year
   old woman has a severe metastasic cancer. She has had several complications and
   is very tired. She has a supportive family and now her husband and herself want
   to know more about her latest complication; a cerebro spinal fluid infection in
   the shunt which delivers chemotherapy to the central nervous system.</profile>
   <narr>Documents should contain information about the infection and its treatment.</narr>
 </topic>
```

Figure 5: Query Examples from CLEF eHealth 2014.

C. Relevance Assessments

Relevance Assessments were collected from professional assessors (not medical experts) using Relevation[12] (White et al., 2009), which is a system for performing relevance judgments for Information Retrieval systems evaluation. The relevance assessment is based on a four-point scale. The relevance grades are (Goeuriot et al., 2014):

- irrelevant,

[12] http://ielab.github.io/relevation/ Retrieved on 03-Mar-2018

- (1) on topic but unreliable,
- (2) relevant,
- (3) highly relevant.

These relevance grades are mapped into a binary scale, as follow:

- Grades (0) and (1) corresponding to the binary grade (0) (irrelevant).
- Grades (2) and (3) corresponding to the binary grade (1) (relevant).

In our experiments, we used the 'title' from the provided fifty-five queries (training and test queries) to validate our system efficiency based on the binary scale relevance assessment.

5.2 Indexing and Experiments

In this section, we discuss the indexing process that we used and the experiments carried out to evaluate our proposed system. We start by explaining the documents pre-processing and indexing settings used to build the inverted index from the document collection. Then, we continue by describing and explaining our experimental runs details with state-of-the-art effectiveness indicators.

A. Documents processing and Indexing

The effectiveness of any information retrieval system (IR system) usually depends on the way and techniques used to process and index the documents in the document collection. In our proposed MIR system, we build a semantically enhanced inverted index that captures the hidden semantic dimensions that are encoded in the text of each indexed medical document in the document collection.

Documents in the CLEF e-Health 2014 document collection are provided as raw web pages that include all the Hypertext Markup Language (HTML) markup and eventually also Cascading Style Sheets (CSS) style definitions and Javascript code. In our experiments, we assume that extracting the text from the web pages will improve the indexing performance, because it removes a great quantity of irrelevant content, like, HTML tags, CSS style, and Javascript codes. To do this, we use Jsoup[13] parser for cleaning and extracting text from raw web pages before starting the indexing process. Jsoup is a Java library for working with HTML. It provides a very convenient Application Programming Interface (API) for extracting and manipulating data, using Document Object Model (DOM), CSS, and jquery-like methods. After text cleaning and extraction, we used a set of NLP techniques to process documents content before building the inverted index. After text extraction, we used natural language processing techniques to process documents content. We first converted all extracted documents text into lower case. Then, we removed stop words based on a predefined list that contains 566 stop words, porter stemmer (Willett, 2006) then used to stem every single term in a document. After that, we add all stemmed terms to our inverted index. If the term related to medical acronyms or abbreviations, (based on UMLS Lexicon), we add all full representations of the term to the inverted index. A sliding window then used to identify and add all compound terms (bi and tri grams) to the index. Then, the UMLS lexicon used another time to check and expand the acronyms and abbreviations available for each bi and trigrams. The identified acronyms for all compound terms added to the inverted index.

We used Oracle 11g database to store our semantically enhanced inverted index. The resulted index size is around 20 GB. To reduce the complexity of the retrieval and ranking process, we perform many calculations that we need during the retrieval process

[13] https://jsoup.org/ Retrieved on 01-Sep-2017

offline and store the calculation values locally with our inverted index. Here we list the calculated information that we store with the inverted index in order to use them during the retrieval process, (we need this information, since we use the $Tf-Idf$ and the vector space model as our information retrieval model):

- The frequency of each single term in each document in the document collection: calculated using $Normalized-tf_{t,d}$ formula (formula 3) discussed in section 4.2.
- The document frequency for each term in the inverted index (df_t): calculated by finding the count of documents containing the term in the document collection.
- The inverse document frequency (Idf_t): calculated using Idf_t formula (formula 2) discussed in section 4.2.
- The term frequency-inverse document frequency (*Tf-Idf)* of each term in each document in the inverted index: calculated using *Tf-Idf* formula (formula 1) discussed in section 4.2.

B. Experiments

This section explains and describes the experiments and the set of runs that we conduct to validate our proposed approach. Experimental instantiation of the proposed system carried out using the CLEF e-Health 2014 dataset to validate the effectiveness of the proposed MIR system for lay people. This validation accomplished through developing a prototype system that applies query reformulation and query terms reweighting and retrieve results from multiple indexed semantic resources to present the final search results.

To evaluate our proposed approach gradually, we carry out five different runs using our system prototype. The first run is the baseline run that we use to measure the effect of using each technique in our proposed approach. The remaining four runs contains the proposed techniques used for improving MIR systems effectiveness. The five runs are described as follow:

- **RUN 1- baseline run**

  In this run, we use a set of basic NLP techniques to process both queries and documents. We use the case folding, punctuation removal, stop words removal and porter stemmer for processing queries at runtime and documents at indexing time. We also use the well-known vector space retrieval model for query-documents matching and ranking.

- **RUN 2**

  In this run, we have re-indexed the document collection by considering medical compound terms (either they are bigrams or trigrams), medical acronyms, and medical abbreviations using the UMLS lexicon as a medical semantic resource. During query processing, we start the query terms classification process by processing both acronyms and abbreviations for query unigrams (*ACt*, *ABt*) - we neglect to process any compound terms during this run-. The main purpose of neglecting compound terms in this run, is that we need to measure the effect of using each processing technique on the system performance and results in effectiveness. The full representations *Et* of both *ACt* and *ABt* added to the reformulated query *Rq*.

The work and results of the following three runs (RUN3, RUN4, and RUN5) are accumulated to the work and results of this RUN. In each following run, we add a newly proposed technique to our system prototype to validate the effectiveness of each single technique on the overall system performance.

- **RUN 3**

  Through this run, we solve the medical compound terms that we neglect in processing users' queries in RUN2. We define a sliding window to identify all compound terms that are either bi or tri grams in the user query. Then we use UMLS lexicon at retrieval time to check if the term (uni, bi and trigram) is a medical acronym or medical abbreviation. Finally, we enrich the user query with the full representations of all identified acronyms and abbreviations for all n-grams in the user query.

- **RUN 4**

  In this run, we use the UMLS Metathesaurus via MetaMap tool and using MRDEF table in the metathesaurus to continue classifying the users' query into its components. We extract medical terms *Mt* (either they are uni, bi or trigrams) and supportive terms *St*, and reweighting query terms based on their classification as we explain in section 4.2 - formula 4.

- **RUN 5**

  This is the last run in our experiments. In this run, we use the UMLS lexicon to extract the synonyms *SYt* of medical terms in the users' query. We add the extracted synonyms to the reformulated query and giving them higher weight over other semantically related concepts *Et* in the reformulated query as we explain in section 4.2 - formula 4.

We will discuss our findings for all runs in the next section (section 5.3). Precision@10 (P@10) evaluation metric was used to evaluate our proposed approach. We choose this metric since it is the most meaningful metric used in web-scale information retrieval systems, and there are multiple experiments done by different authors using the same dataset we used (CLEF e-heath 2014) calculating this measure which allow us to compare our findings with others. P@10 corresponds to the number of relevant results on the first 10 search results retrieved by retrieval system. The next formula used for P@10 calculation (Note: the maximum no. of results returned by our prototype is 10 documents - top 10 high ranked documents):

$$P@10 = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \qquad (6)$$

Where the documents retrieved by our system compared with relevant assessment provided with CLEF e-health 2014 dataset to distinguish relevant and non-relevant documents.

5.3 Experimental Results and Discussion

In this section, first, we conduct our experiments to evaluate the effectiveness of reformulating user's queries and reweighting query terms by comparing the precision effectiveness indicator of each prototype run with the precision of the baseline run. On the other hand, in order to validate the effectiveness of our proposed system we compare the precision indicator value produced by our final system prototype with three state-of-the-art systems proposed recently in MIR domain.

A. Evaluating our Proposed MIR System Runs with the Baseline Run

In this section, we compare the P@10 values of our different proposed runs with the P@10 values of the system baseline run and discuss the comparison results. The different system runs described in detail in section 5.2 and a brief overview of those runs summarized in table 3.

Table 3 : Experiments setup for our 5 system runs

RUN	Experiment methods
1	Baseline (simple inverted index with cosine similarity retrieval model)
2	Semantically enhanced inverted index with acronyms and abbreviation processing using UMLS lexicon
3	Experiment methods used in run 2 in addition to compound terms processing
4	Experiment methods used in run 3 in addition to medical terms processing using UMLS metathesaurus and query terms reweighting technique
5	Experiment methods used in run 4 in addition to medical synonyms enrichment using UMLS lexicon

To evaluate our proposed system step by step, we use the following figures (figure 6, figure 7, figure 8 and figure 9) to compare the system runs with a baseline run at P@10 effectiveness indicator. Figure 6 shows the comparison results for run 2 with the baseline run.

For RUN 2 as figure 6 shows, among the total 55 test and training queries, 1 query (qtest2014.27) have worsen the performance than the baseline run; 3 queries (qtest2014.3, qtest2014.433 and qtest2014.47) have the best performance; the other 51 queries have performances equal to the baseline run. The main reason for the bad performance of the qtest2014.27 is that this query contains an acronym that has multiple full representation forms which leads retrieving not related documents to the query context.

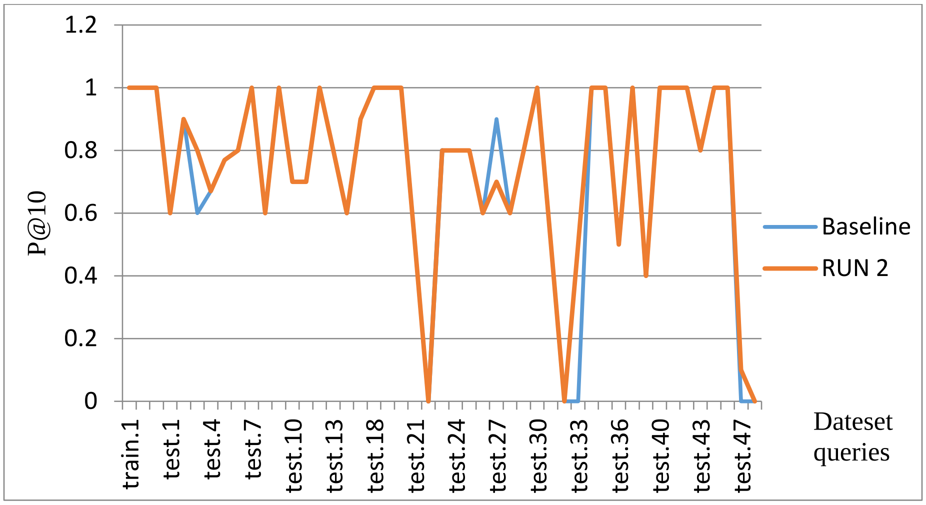


Figure 6 : Baseline run VS RUN 2 at P@10

Figure 7 shows the comparison results for run 3 with the baseline run. As we can notice, there are 2 queries (qtest2014.24 and qtest2014.27) have worsen the performance than the baseline run; and 4 queries (qtest2014.3, qtest2014.15 ,qtest2014.433 and qtest2014.47) have the best performance; the other 49 queries have performances equal to the baseline run. The main reason for the worse performance of the qtest2014.24 is that this query contains bi and trigrams that can related to many contexts rather than the query context only. For query qtest2014.27, its worse results are accumulated from the previous run (RUN 2).

The most significant improvement of the proposed system effectiveness over the baseline run has been achieved in run 4. As we can see in figure 8, 11 among 55 queries have the best performance than the baseline run. We can infer by this results improvement that the terms reweighting techniques which based on query terms classifications; is the best proposed techniques that help in improving the MIR systems effectiveness. The proposed query terms classification process that exploits UMLS metathesaurus via MetaMap tool and via MRDEF table was the best choice for classifying the query terms as depicted in the results of run 4.

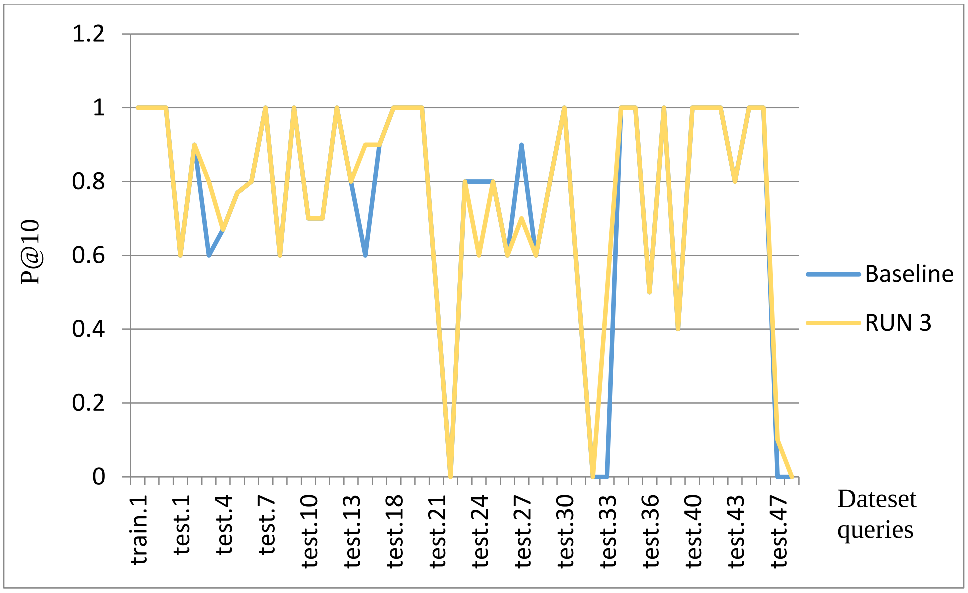


Figure 7: Baseline run VS RUN 3 at P@10

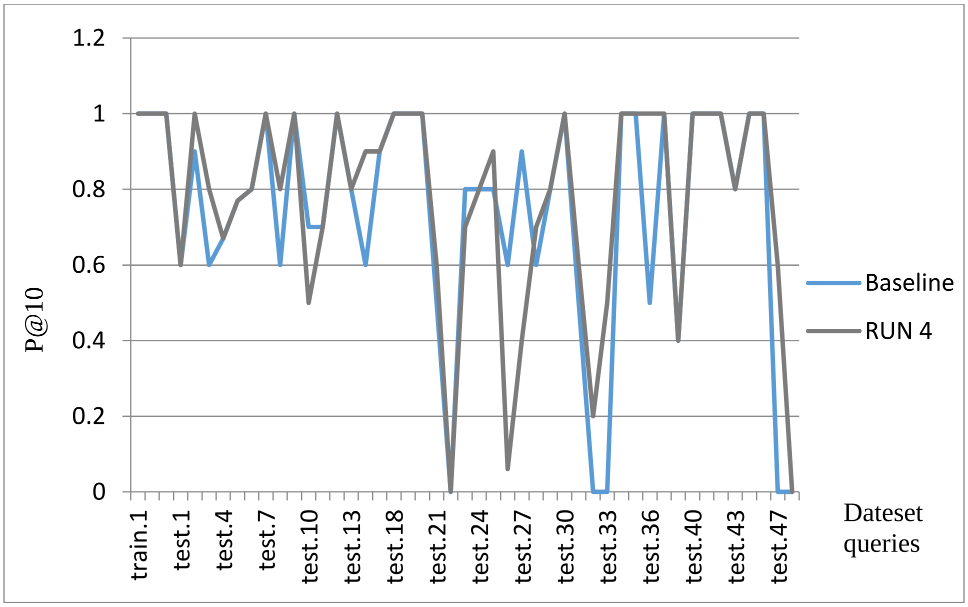


Figure 8: Baseline run VS RUN 4 at P@10

The last module in our proposed query reformulation component is the query enrichment with medical synonyms that related to query medical terms. In run 5, we validate the results of this module as an accumulated step in our proposed system. Figure 9 showed the comparison between the P@10 of the run 5 and the baseline run. As we can notice from figure 9, another improvement on the system effectiveness also achieved by run 5. By comparing the results of run 5, we compare the final prototype of our proposed system. A significant improvement over the baseline system is achieved by our proposed system. But, we still need more improvements to solve the queries that have a worse effect such as qtest2014.24 and qtest2014.27.

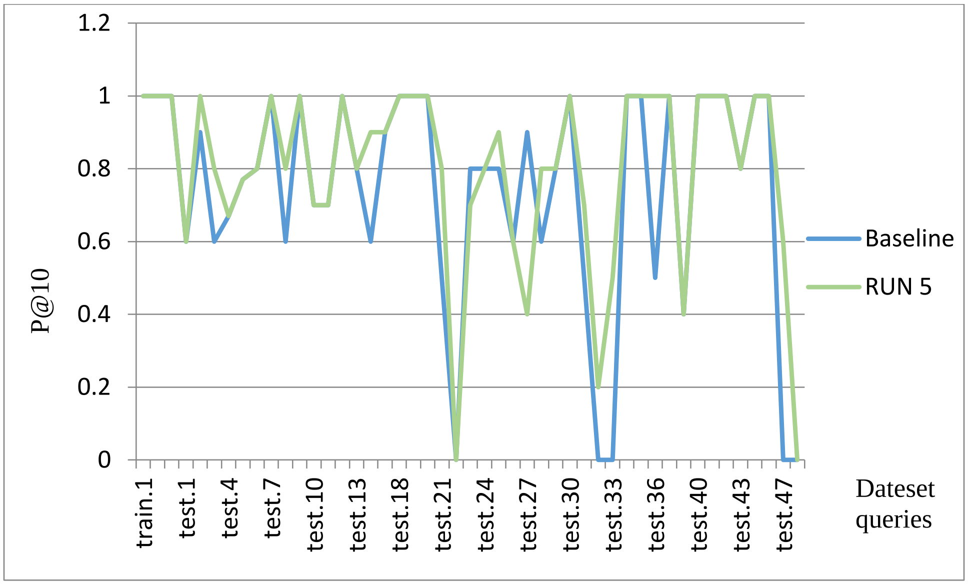


Figure 9: Baseline run VS RUN 5 at P@10

Figure 10 provides a comprehensive comparison for all systems' runs with the baseline run based on the achieved precision values.

The overall precision results of our different proposed system runs are summarized in Table 4. These values represent the proposed system performance after testing all training and testing queries provided in CLEF e-health 2014 dataset (5 training queries and 50 test queries).

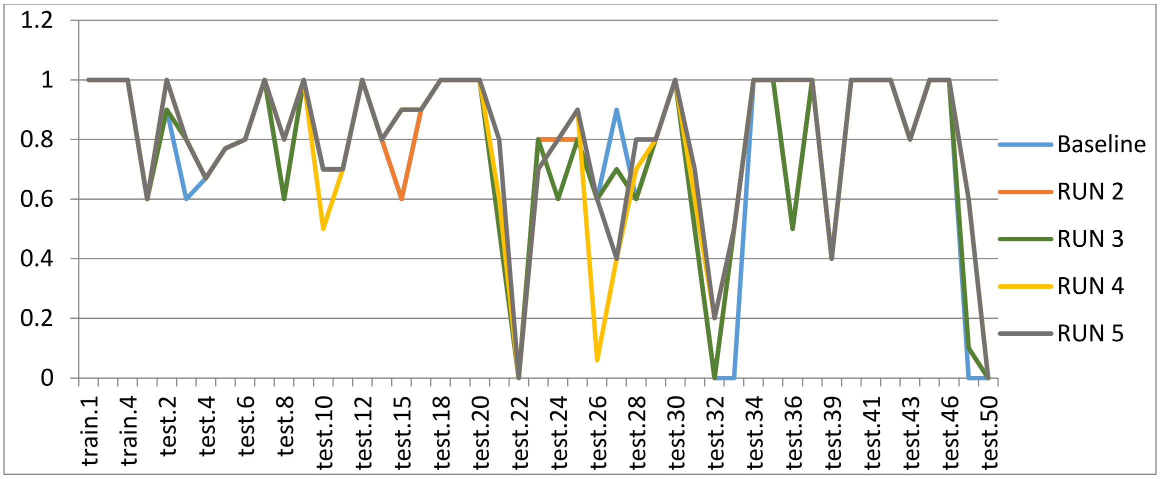


Figure 10: Comprehensive comparison for all systems' runs with the baseline run

Table 4: MedIR system performance over system runs

Runs	**Baseline**	**Run 2**	**Run 3**	**Run 4**	**Run 5**
P@10	0.7232	0.7500	0.7520	0.7723	**0.7966**

The experimental results show that the methodologies proposed in our retrieval model improved the precision of our baseline run (traditional bag of words retrieval model) by around 0.0734. Based on the previous results depicted in table 4, we can notice that the methodologies proposed in our medical information retrieval model improved the precision of our baseline run (traditional bag of words retrieval model) by around 0.0734. Our results indicate that exploiting integrated medical semantic resources for enriching laypeople queries in the medical domain, and also reweighting query terms based on their classification improve the effectiveness of MIR systems.

B. Evaluating the MedIR System effectiveness with other state-of-the-art Systems

In this section, we compare the results produced by our final system prototype in run 5 to those produced by the best three CLEF participant teams who used the CLEF e-Health 2014 dataset for evaluating their proposed systems (Goeuriot et al., 2014):

- GRIUM (Shen et al., 2014) in their best run (EN-Run.5) : their methodology and results discussed in chapter 2.
- SNUMEDINFO (Choi and Choi, 2014) in their best run (EN-Run.2): their methodology and results discussed in chapter 2.
- KISTI (Oh and Jung, 2014) in their best run (EN-Run.2): their methodology and results discussed in chapter 2.

Table 5: Comparison between the effectiveness of the MedIR system and three state-of-the-art MIR systems

P@10	MedIR	GRIUM	SNUMEDINFO	KISTI
Baseline run	0.7211	0.7180	**0.7380**	0.7300
Best Run	**.7945**	0.7560	0.7540	0.7400

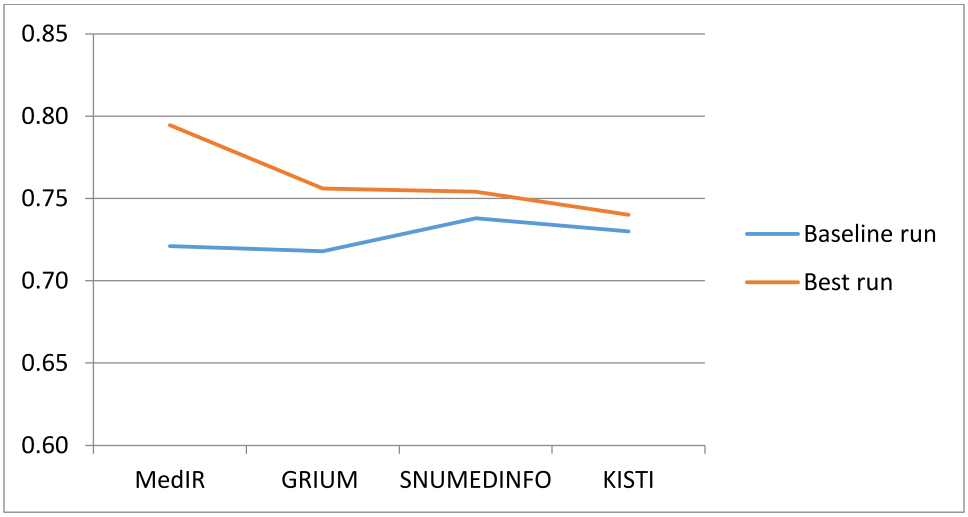


Figure 11: Comparison between the effectiveness of the MedIR system and three state-of-the-art MIR systems

As table 5 and figure 11 show, our proposed MedIR system achieves more improvement than others. regarding the P@10 effectiveness indicator. One of the basic factors that may have improved our results, is the use of a semantically enhanced inverted index that solving compound terms and both medical acronyms and abbreviations using external medical semantic resources during the indexing process. Where the other

systems mentioned here used basic existing Indexing and retrieval systems (Indri and Lucene). GRIUM authors proposed a retrieval model using a bag of concepts instead of a traditional bag of words retrieval model. They used MetaMap tool for extracting medical concepts exist in the user query to be considered in their query-document matching process and ignored all other supportive terms in the original query. The main drawbacks of GRIUM proposed system is that they neglect all query terms that are not identified as medical concepts by MetaMap, which leads to possibility for ignoring some medical concepts such as medical acronyms and abbreviations, and this explains the slight improvements in their system precision (0.7560) over the baseline run where the precision was (0.7180). In SNUMEDINFO, authors used a simple inverted index (compound terms not included) with UMLS metathesaurus for query expansion. The main drawbacks in SNUMEDINFO that reduces the effect of their proposed approach are: 1) They neglect the compound terms that may occur in both of the document collection or in the user query. 2) They also neglect the semantic type (synonym, mapped to, part of, .. etc) of concepts extracted from UMLS metathesaurus and giving all extracted concepts the same importance as the original query terms. 3) They didn't solve acronyms and abbreviations in the documents and queries. KISTI proposed system achieved little improvement in their system precision over the baseline and compared with other MIR systems. The main reason for this is that the authors used the related <discharge summary> information provided with each query in the dataset as their expansion resource instead of using external medical semantic resources, while the discharge summary reports can't be considered as a trusted enriched medical semantic resource that can be used as effective expansion resource for medical queries. In our proposed system, we overcome the drawbacks of the existing MIR systems to achieve better performance than they do.

5.4 Summary

In this chapter, we discussed the experiments that we have conducted to validate the efficiency and the effectiveness of the proposed MIR system. In addition, we have compared the produced results by our system with three state-of-the-art systems.

During this chapter, we used two evaluation techniques to validate our experimental results. First, we evaluate the effectiveness of each proposed module by comparing the precision improvement achieved by integrating such module to the system prototype with the precision of the baseline run. The conducted experiments using the exploited medical semantic resources and proposed terms reweighting techniques; assists in achieving higher precision results than using traditional information retrieval techniques implemented in the baseline system.

In the second evaluation technique, we compare the precision value achieved by integrating all system modules we propose into one system prototype with three state-of-the-art systems proposed recently in the medical domain and evaluated using the same evaluation dataset we used in our experiments. The conducted comparison showed that employing our integrated medical semantic resources and terms reweighting techniques;

have led to significant enhancements in the results effectiveness and system precision. These enhancements achieved due to enriching the user query with related specialized medical terms that help in best describe lay user's needs, and also due to give higher weight for medical terms in the user query over other existed supportive terms.

Chapter 6 - Conclusions and Future Work

This chapter summarizes our proposed approach for building MIR system, discusses its findings and contributions that we achieved from building the proposed system. In addition, it outlines the future extensions for the current version of our proposed system. The chapter is divided into two sections. Section 6.1 presents a discussion of the contribution of our research work and highlights the techniques/approaches that we utilize in the proposed system. Section 6.2 discusses the future works and the other challenges that we plan to tackle in the future system updates.

6.1 Conclusions

In this thesis, we have proposed a medical information retrieval system that aims to fill the semantic gap and go beyond keywords matching between the lay user's queries and their related documents in the medical domain. By doing this, Our proposed approach improves the retrieved results for laypeople when they searching the web in the medical domain. To meet this goal, we propose a medical information retrieval system (MedIR) that improves the effectiveness of the currently available MIR systems by:

1. Constructing a semantically enhanced Inverted Index for indexing medical documents. The main aim of this inverted index is to capture the hidden semantic dimensions that are encoded in the text of each medical document. We used the UMLS lexicon together with a set of NLP techniques - especially n-gram tokenization- to build this enhanced inverted index.
2. Reformulating the given users’ queries through enriching them with semantically related medical concepts such as, the full representations of the medical acronyms and abbreviations, and also the synonyms of all medical terms in the users' query.

For the expansion process, We used two external medical semantic resources: UMLS lexicon and UMLS metathesaurus. By reformulating he users' queries we minimized the gap between laypeople queries and the terminology of the medical documents.

3. Classifying the users' query terms into medical and supportive terms using an integrated external medical semantic resources and assigning a higher weight for the medical terms over other supportive terms.

In order to evaluate the effectiveness of the proposed system, we used CLEF e-Health 2014 dataset that is constructed for evaluating MIR systems. First, we compared our system performance with the basic information retrieval system that we built to compare the improvement of our MedIR system with the traditional information retrieval systems. Moreover, in order to validate the effectiveness of our MedIR system, we use the well known precision indicator, to compare our results with three recently proposed MIR systems. The conducted comparison showed that our MedIR system is more efficient than the others with around five to seven percent.

6.2 Challenges and Future work

Although the conducted experiments showed promising results, there are other potential improvements to the techniques presented in this research work. Below we discuss these improvements and outline proposals on how to achieve them in our future work:

A. In the Query reformulation component, we used integrated medical semantic resources to enrich all user's queries with semantically related medical concepts. However, we don't know if this enrichment will improve the results of all user's queries. Accordingly, we plan to improve our system by implementing a selective query expansion model that will predict the effect of the reformulation process on each single query, to decide if reformulating a query will improve the search results for such query or not to decide if we will reformulate it.

B. In the medical acronyms and abbreviations recognition and enrichment module, we used the UMLS lexicon for identifying and expanding medical acronyms and abbreviations in the original query. All the expected expansion and full representations of acronyms and abbreviations added to the original query. However, our proposed module didn’t take into consideration the context of the original query during the enrichment process. In order to increase the effectiveness of the enrichment process, we aim to classify the enrichment possibilities and adding only concepts that cope with the users' query context.

C. In our proposed approach, we used UMLS lexicon and UMLS metathesaurus to reformulate all user's queries. But, we don't take into consideration that the use of some medical semantic resources will improve the effectiveness of some queries and degrade the effectiveness of other queries. We plan to use more semantic resources for query reformulation such as MeSH (Lipscomb, 2000), ICD-10 standard (Coustasse and Paul III, 2013) and Pubmed (Malhotra et al., 2015). Before reformulating any user query, we plan to use a prediction technique to measure the efficiency of using all medical semantic resources and then choose the best resource for each query.

References

ABDERRAHIM, M. E. A. 2014. Concept based vs. pseudo relevance feedback performance evaluation for information retrieval system. *arXiv preprint arXiv:1403.4362*.

ABDOU, S. & SAVOY, J. 2008. Searching in Medline: Query expansion and manual indexing evaluation. *Information Processing & Management,* 44**,** 781-789.

ARONSON, A. R. & LANG, F.-M. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association,* 17**,** 229-236.

BODENREIDER, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research,* 32**,** D267-D270.

CARPINETO, C., DE MORI, R., ROMANO, G. & BIGI, B. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS),* 19**,** 1-27.

CHOI, S. & CHOI, J. 2014. Exploring Effective Information Retrieval Technique for the Medical Web Documents: SNUMedinfo at CLEFeHealth2014 Task 3. CLEF (Working Notes). 167-175.

CHRISTOPHER, D. M., PRABHAKAR, R. & HINRICH, S. 2008. *Introduction to information retrieval*.

CONSORTIUM, G. O. 2011. The gene ontology: enhancements for 2011. *Nucleic acids research,* 40**,** D559-D564.

COUSTASSE, A. & PAUL III, D. P. 2013. Adoption of the ICD-10 standard in the United States: The time is now. *The health care manager,* 32**,** 260-267.

DÍAZ-GALIANO, M. C., GARCÍA-CUMBRERAS, M., MARTÍN-VALDIVIA, M. T., MONTEJO-RÁEZ, A. & URENA-LÓPEZ, L. 2007. Integrating mesh ontology to improve medical information retrieval. Workshop of the Cross-Language Evaluation Forum for European Languages. Springer, 601-606.

FOX, S. & DUGGAN, M. 2013. Health online 2013. *Health,* 2013**,** 1-55.

FREITAS-JUNIOR, H. R., RIBEIRO-NETO, B., VALE, R. F., LAENDER, A. H. & LIMA, L. R. 2006. Categorization-driven cross-language retrieval of medical information. *Journal of the American Society for Information Science and Technology,* 57**,** 501-510.

GOEURIOT, L., JONES, G. J., KELLY, L., MÜLLER, H. & ZOBEL, J. 2016. Medical information retrieval: introduction to the special issue. *Information Retrieval Journal,* 19**,** 1-5.

GOEURIOT, L., KELLY, L., LI, W., PALOTTI, J., PECINA, P., ZUCCON, G., HANBURY, A., JONES, G. J. & MUELLER, H. 2014. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. Proceedings of CLEF 2014.

HANNA, J., JOSEPH, E., BROCHHAUSEN, M. & HOGAN, W. R. 2013. Building a drug ontology based on RxNorm and other sources. *Journal of biomedical semantics,* 4**,** 44.

HERSH, W., BUCKLEY, C., LEONE, T. & HICKAM, D. 1994. OHSUMED: an interactive retrieval evaluation and new large test collection for research. SIGIR'94. Springer, 192-201.

HERSH, W. R., BHUPATIRAJU, R. T., ROSS, L., ROBERTS, P., COHEN, A. M. & KRAEMER, D. F. 2006. Enhancing access to the Bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration,* 1**,** 3.

HLIAOUTAKIS, A., VARELAS, G., VOUTSAKIS, E., PETRAKIS, E. G. & MILIOS, E. 2006. Information retrieval by semantic similarity. *International journal on semantic Web and information systems (IJSWIS),* 2**,** 5.73-5

HUANG, Y., LOWE, H. J., KLEIN, D. & CUCINA, R. J. 2005. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. *Journal of the American Medical Informatics Association,* 12**,** 275-285.

KEITH, A. & BURRIDGE, K. 1991. Euphemism and dysphemism: language used as shield and weapon. Oxford: Oxford University Press.

KIBBE, W. A., ARZE, C., FELIX, V., MITRAKA, E., BOLTON, E., FU, G., MUNGALL ,C. J., BINDER, J. X., MALONE, J. & VASANT, D. 2014. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research,* 43**,** D1071-D1078.

KIM, E. H.-J., OH, J. S & .SONG, M. 2015. Exploring Context-Sensitive Query Reformulation in a Biomedical Digital Library. *In:* ALLEN, B. R., HUNTER, J. & ZENG, L. M. (eds.) *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, Seoul, Korea, December 9-12, 2015. Proceedings.* Cham: Springer International Publishing.

KOOPMAN, B., ZUCCON, G., BRUZA, P., SITBON, L. & LAWLEY, M. 2016. Information retrieval as semantic inference: A graph inference model applied to medical search. *Information Retrieval Journal,* 19**,** 6-37.

LARSON, R. R. 2010. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology,* 61**,** 852-853.

LEE, D., CORNET, R., LAU, F. & DE KEIZER, N. 2013. A survey of SNOMED CT implementations. *Journal of biomedical informatics,* 46**,** 87-96.

LIPSCOMB, C. E. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association,* 88**,** 265.

LOSEE, R. M. 2001. Term dependence: A basis for Luhn and Zipf models. *Journal of the American Society for Information Science and Technology,* 52**,** 1019-1025.

LU, F. 2015. Employing Query Expansion Models to help Patients Diagnose Themselves. CLEF (Working Notes).

MALHOTRA, A., GÜNDEL, M., RAJPUT, A. M., MEVISSEN, H.-T ,.SAIZ, A., PASTOR, X., LOZANO-RUBI, R., MARTINEZ-LAPSICINA, E. H., ZUBIZARRETA, I. & MUELLER, B. 2015. Knowledge retrieval from PubMed abstracts and electronic medical records with the Multiple Sclerosis Ontology. *PloS one,* 10**,** e0116718.

MOGOTSI, I. C. 2 .010Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval. *Information Retrieval,* 13**,** 192-195.

OH, H.-S. & JUNG, Y. A. 2014. Multiple-stage Approach to Re-ranking Clinical Documents. CLEF (Working Notes).219-210 .

PALOTTI, J., HANBURY, A. & MÜLLER, H. 2014. Exploiting health related features to infer user expertise in the medical domain. Web Search Click Data workshop at WSCM, New York City, NY, USA.

PALOTTI, J., HANBURY, A., MÜLLER, H. & KAHN, C. E. 2016 .How users search and what they search for in the medical domain. *Information Retrieval Journal,* 19**,** 189-224.

RIEH, S. Y. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management,* 42**,** 751-768.

RUTHVEN, I. & KELLY, D. 2011. *Interactive information seeking, behaviour and retrieval,* Facet Publishing.

SCHWARTZ, A. S. & HEARST, M. A. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. *Biocomputing 2003.* World Scientific.

SHEN, W., NIE, J.-Y., LIU, X. & LIUI, X. 2014. An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@ CLEF2014eHealthTask 3. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

SOLDAINI, L., YATES, A., YOM-TOV, E., FRIEDER, O. & GOHARIAN, N. 2016. Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal,* 19**,** 149-173.

STANTON, I., IEONG, S. & MISHRA, N. 2014. Circumlocution in diagnostic medical queries. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 133-142.

SUSANNAH FOX, M. D. 2013. *Susannah Fox and Maeve Duggan* [Online]. Available: http://www.pewinternet.org/2013/15/01/health-online-2013./

SZOLOVITS, P. 2003. Adding a medical lexicon to an English parser. AMIA annual symposium proceedings. American Medical Informatics Association, 639.

THESPRASITH, O. & JARUSKULCHAI, C. 2014. CSKU GPRF-QE for Medical Topic Web Retrieval. CLEF (Working Notes). Citeseer, 260-268.

TIBI, O., THUMA, E. & MOSWEUNYANE, G. 2017. Selective collection enrichment in user-centred health information retrieval. Next Generation Computing Applications (NextComp), 2017 1st International Conference. IEEE, 175-181.

VOORHEES, E. M. & HARMAN, D. 2000. Overview of the sixth text retrieval conference (TREC-6). *Inf. Process. Manage.,* 36**,** 3-35.

VREEMAN, D. J., MCDONALD, C. J. & HUFF, S. M. 2010. LOINC®: a universal catalogue of individual clinical observations and uniform representation of enumerated collections. *International journal of functional informatics and personalised medicine,* 3**,** 273-291.

WHITE, R. W., DUMAIS, S. T. & TEEVAN, J. 2009. Characterizing the influence of domain expertise on web search behavior. Proceedings of the second ACM international conference on web search and data mining. ACM, 132-141.

WILLETT, P. 2006. The Porter stemming algorithm: then and now. *Program,* 40**,** 219-223.

XU, J. & CROFT, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS),* 18**,** 79-112.

YOO, S. & CHOI, J. 2010. On the query reformulation technique for effective MEDLINE document retrieval. *Journal of biomedical informatics,* 43**,** 686-693.

ZGHAL, H. B. & MORENO, A. 2014. A system for information retrieval in a medical digital library based on modular ontologies and query reformulation. *Multimedia tools and applications,* 72**,** 2393-2412.

ZICKUHR, K. *Who's not online and why*. http://www.pewinternet.org/2013/09/25/whos-not-online-and-why-2/ [Accessed 10-Aug-2018.

حاليا - والتي استخدمت نفس مجموعة البيانات لتقييم الانظمة المقترحة - باستخدام مؤشر الدقة المعروف. لقد اظهرت التجارب التي تم اجراؤها باستخدام مجموعة البيانات المذكورة سابقا ان النظام المقترح حقق مؤشر دقة بقيمة 0.7945، بينما كانت قيم مؤشر الدقة لأنظمة استرجاع المعلومات الطبية المماثلة التي تم اجراء المقارنة بها كالتالي (0.7400 ,0.7540 ,0.7560). ان قيم مؤشر الدقة السابقة تبين ان جودة النتائج المسترجعة باستخدام النظام المقترح قد تم تحسينها بحوالي خمسة الى سبعة بالمئة بالمقارنة مع الانظمة الاخرى المشابهة.

الملخص باللغة العربية

اعادة تشكيل وصياغة جمل البحث في مجال استرجاع المعلومات الطبية

في السنوات الأخيرة أصبحت محركات البحث أدوات لا غنى عنها لمساعدة مستخدميها في العثور على المعلومات في غالبية اهتماماتهم. ويمكن تصنيف محركات البحث هذه إلى فئتين رئيسيتين؛ محركات البحث العامة (والتي تعرف بمحركات البحث الأفقية) ومحركات البحث المتخصصة (والتي تعرف بمحركات البحث العمودية. يستغل كلا النوعين من محركات البحث خوارزميات وتقنيات مختلفة لفهرسة الوثائق على الويب أولاً ومن ثم استرجاع النتائج الأكثر ملاءمة لجملة البحث والتي تفي باحتياجات المستخدمين من المعلومات. في هذا البحث ، سيكون تركيزنا على الفئة الثانية من محركات البحث - أي البحث في مجال محدد - وتحديدًا في المجال الطبي. عند استخدام محركات البحث الطبية ، تعتمد دقة النتائج المسترجعة بشكل رئيسي على عاملين رئيسيين. أولاً ، حاجة المستخدمين (سواء كانوا أشخاصًا عاديين أو محترفين طبيين) إلى تقديم مفردات تصف احتياجاتهم من المعلومات على أفضل وجه. ثانياً ، استناد جودة النتائج المسترجعة إلى حد كبير على فعالية تقنيات البحث ومصادر المعرفة الطبية المستخدمة بواسطة محركات البحث هذه.

تم اقتراح العديد من الأنظمة والمنهجيات لمعالجة المشاكل المتعلقة بكل من هذه العوامل بشكل مستقل ؛ مثل استخدام الكلمات العامة للتعبير عن احتياجات الاشخاص العاديين بدلا من استخدام المصطلحات الطبية الخاصة، وكذلك استخدام الموارد المعرفية التي تحتوي على عدد محدود من المفاهيم الطبية مما يؤدي الى وجود فجوة بين جمل البحث للأشخاص العاديين وبين الوثائق والملفات الطبية ذات العلاقة. ولكن ، القليل من الاهتمام من قبل الباحثين تم اعطاؤه لمعالجة المشاكل المتعلقة بكلا العاملين في آن واحد. نهدف في هذا البحث إلى دراسة تأثير استغلال الموارد الدلالية الطبية وتقنيات استرجاع المعلومات في 1) إعادة صياغة استفسارات المستخدمين من خلال إثرائها بمفاهيم ذات دلالات لغوية مختلفة متعلقة باستفساراتهم و 2) فهرسة المستندات في المجال الطبي لتحسين جودة نتائج البحث المسترجعة. لإثبات فعالية النظام المقترح ، أجرينا العديد من التجارب باستخدام مجموعة بيانات معتمدة ومخصصة لغرض تقييم انظمة استرجاع المعلومات الطبية وهذه المجموعه هي (CLEF eHealth2014)، بالإضافة إلى ذلك، قمنا بتقييم كفاءة وفعالية النظام المقترح (MedIR) بالمقارنة مع ثلاثة من أنظمة استرجاع المعلومات الطبية المتوفرة