January 2019

Utilizing Knowledge Bases In Information Retrieval For Clinical Decision Support And Precision Medicine

Saeid Balaneshinkordan
Wayne State University, balaneshin.saeid@gmail.com

UTILIZING KNOWLEDGE BASES IN INFORMATION RETRIEVAL FOR

CLINICAL DECISION SUPPORT AND PRECISION MEDICINE

by

SAEID BALANESHINKORDAN

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: COMPUTER SCIENCE

Approved By:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

To My Wife Elaheh

ACKNOWLEDGEMENTS

I would like to express my warmest gratitude to my advisor Dr. Alexander Kotov who believed in me and encouraged me to continue learning, working hard, and taking new challenges. I am very grateful for his encouragement and support and all his comments and discussions. I would like to sincerely thank my PhD dissertation committee members Dr. Dongxiao Zhu, Dr. Zichun Zhong, and Dr. Matthew Nokleby for their valuable feedback on my work, their tremendous support and useful suggestions. I will forever be thankful to the alumni and members of the Textual Data Analytics Laboratory (TEANA) group. Many thanks to my friends and lab mates Fedor Nikolaev, Mehedi Hassan and Diana Diaz for their insightful discussions and feedbacks that helped improve this work. Also, I would like to thank Dr. Ming Dong and his students Shixing Chen, Haotian Xu, and Hajar Emami for being great labmates. I would like to express my gratitude to the Department of Computer Science and its chair Dr. Loren Schwiebert for their help and support. Many thanks and acknowledgments go to my parents for all their help, support and the chances they have given me over the years. Above all, I would like to thank my wife Elaheh for standing beside me throughout my graduate studies and writing this dissertation. I am sincerely grateful for her many contributions to my work and my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1. Medical Information Retrieval

Medical Information retrieval (IR) can be considered as one of the most challenging information retrieval tasks, and just like any other medical task is of the highest priority. Although medical Information retrieval is a branch in information retrieval field, it has roots in other fields such as Clinical Decision Support (CDS) and Precision Medicine (PM). This task has been formulated for a variety of purposes, such as gnomic information retrieval [78], and its proposed methods can be categorized into three main classes. Those based on statistical methods, those based on knowledge bases and those based on a combination of them [106].

We consider medical IR scenarios for CDS and PM in which a query describes a patient case that can consist of multiple components including several sentences, name of disease, type of query, genes mutated, etc. In these medical IR scenarios, given a query provided by a clinical practitioner, we aim at finding relevant articles in medical literature that would support her in her decision-making process. Accurately answering information needs in CDS and PM tasks requires utilizing a variety of resources such as knowledge bases to capture explicit and latent query concepts and to determine their relative importance.

PM [28] is a recent initiative that aims at personalizing healthcare by taking into account variability across different patients at the physiological and molecular levels. Successful realization of this initiative requires both significant advances in biomedical research, such as methods for the accurate assessment of the risk of healthy individuals developing a disease and selection of the optimal therapy for patients with a particular disease, and adoption of these scientific advances in clinical practice. By providing clinicians with supporting information in the

form of scientific articles relevant to a given description of a patient case, IR systems for CDS are a crucial link between scientific advances and clinicians at the point of care. Therefore, the development of methods and models for biomedical IR [90, 113, 37] that can provide clinicians with fast and reliable access to relevant biomedical publications and thus facilitate the selection of an optimal treatment for each individual patient from a large number of options is an important component of PM.

Recently, an **IR task for CDS** was proposed in a special track on Clinical Decision Support in the Text REtrieval Conference (TREC-CDS) [95, 91, 89]. In this task, the queries typically correspond to complex information needs, which involve a large number of concepts of different types from different query fields such as patient demographics, symptoms of a disease or test results. For example, the query "*A 4-year-old girl presents with persistent fever for the past week. The parents report a spike at 104° F. The parents brought the child to the emergency room when they noticed erythematous rash on the girl's trunk. Physical examination reveals strawberry red tongue, red and cracked lips, and swollen red hands. The whites of both eyes are red with no discharge.*", includes the query concepts that indicate the age and gender of a patient, describes several symptoms, such as erythematous rash, and test results, such as revealed swollen red hands and strawberry red tongue, as well as indicates a possible diagnosis, such as strawberry tongue (also known as Kawasaki disease). Although such queries are fairly long, only a fraction of concepts corresponding to an information need underlying those queries are directly mentioned in them, such as the concept strawberry tongue in the query above (i.e., explicit concepts), while many other concepts representing the same information need do not occur in the queries themselves, but can be found in other resources such as knowledge bases and pseudo-relevance

feedback (PRF) documents (i.e., latent concepts). For example, the concept Kawasaki disease that is not explicitly mentioned in the above query, can be found in UMLS knowledge base as one of the related concepts to the concept strawberry tongue and also in an article with the PubMed unique identifier (PMID) 3625593 (i.e., [33]), which is among the top retrieved documents for this query. Some of the information that exist in the raw query may not directly be used in a textual information retrieval system to retrieve clinical relevant documents. For example, gene mutation information that is described by a number of symbols for the type of gene and its mutation cannot be used as a keyword query because of their sparsities in a collection of textual documents such as PubMed. Therefore, for this type of information in the query, we need to mainly rely on its relevant concepts that we may obtain from a combination of knowledge bases.

As in the general case of IR systems for CDS, the goal of **IR tasks for CDS in PM** [28] is to help healthcare providers find documents that are relevant to a patient case in an archive of biomedical articles. For example, a clinician may pose a query that includes information about the cancer type, patient age, gender and other factors regarding the patient case, such as gene mutations. In general, queries posed to IR systems for CDS in PM have three distinct properties. First, these queries are significantly shorter than medical case descriptions. Therefore, the proposed method is focused on effective query expansion rather than on information extraction and concept weighting. Second, these queries are structured with the fields of queries of differing importance. Third, these queries contain both textual and non-textual information. Specifically, they typically include genetic variant data (e.g., mutations in patient genes characterized by the gene name, such as “PIK3CA”, and amino acid (AA) position codes within the mutated gene, such as “E545K”). Genetic variants play an important role in personalizing treatment because they can

cause complex diseases, such as cancers, that share a similar set of symptoms to respond differently to the same treatment [8]. Therefore, the proposed method is focused on effectively incorporating gene mutation information into biomedical article retrieval.

Recently proposed approaches to **identify and weight query concepts** are either based only on semantics [55, 101, 44, 103, 128] or are purely statistical [104, 69, 15, 17, 49, 70, 127]. Each of these two types of approaches are able to identify only certain types of concepts. For example, [55] identifies and utilizes only the concepts from the Unified Medical Language System (UMLS) that are extracted using the MetaMap tool [5] from PRF documents. Single-word and multi-word statistical concepts from the query and single-word concepts from PRF documents have been shown to be effective for ad-hoc retrieval in [70, 17]. A bag-of-words retrieval model utilizing medical concepts from PRF documents for query expansion was proposed in [101]. Choi et al. [27] proposed a method to represent multi-word UMLS concepts using sequential dependencies between their words.

While previous work on general and domain-specific IR has focused on identification of the key statistical concepts in verbose queries [15, 16, 17], latent query concepts in external resources ([56, 103, 128, 129]) and the top-retrieved (PRF) documents [17, 70] individually, **query transformation** methods that use both explicit concepts from the query and latent concepts from diverse sources, such as external resources and PRF documents, has been less investigated. For example, Latent Concept Expansion (LCE) [70] and Parameterized Query Expansion (PQE) [17] methods use only unigrams from the top-retrieved documents as latent concepts, while [48] uses only unigrams from structured knowledge bases as latent concepts for query expansion.

1.2. Knowledge bases in Medical Information Retrieval

Vocabulary mismatch and underspecified queries, which contain only a fraction of concepts that represent the information need (henceforth referred to as explicit concepts), are the two most common reasons for inaccurate and incomplete search results. Knowledge-bases can improve the retrieval quality by providing a possibility to fill the information gap between users and the machine. The knowledge that knowledge bases provides, may not be achieved from other resources like top-ranked documents. One of the ways to use knowledge-bases for information retrieval task is to extract the most relevant concepts from the list of related concepts in the knowledge base. Semantic types that are also provided by knowledge bases can be also used to narrow down the concepts into concepts that semantically are more related to the query.

Synonyms of explicit concepts, as well as other concepts that are relevant to the information need, but are not explicitly mentioned in the query (henceforth referred to as latent concepts), can be extracted either from the top retrieved documents [17, 25, 70, 51] or from external knowledge repositories [30, 48, 117, 118, 119], such as knowledge bases and semantic networks, and added to a query through the query expansion process. Knowledge bases and knowledge graphs can be very effective for entity-bearing queries and are primarily utilized by first linking queries to entities in a knowledge graph [39, 87] and then enriching the query with elements of textual entity representations, including entity names, the names of related entities, categories and structured attributes [30, 118]. Leveraging general-purpose or domain-specific semantic networks or concept graphs, in which the nodes correspond to words or phrases and the typed edges designate semantic relationships between them, is an alternative approach to

query expansion that we focus on in this work. Such approach is applicable to any query, since it does not require query to contain entities that can be linked to a knowledge base.

1.3. Concept graphs

Concept graphs are used in domain-specific and general information retrieval systems to identify latent concepts in query and thus to fill the information gap between users and the retrieval systems [101, 55]. In domain-specific IR systems such as in medical IR systems, the source of these concepts can be either domain-specific such as Unified Medical Language System (UMLS) or general-purpose such as ConceptNet. In UMLS, each concept may correspond to one or multiple semantic types, which provide semantic information about UMLS concepts. In [55], the authors have proposed that the semantic types of concepts can be used to extract concepts from the concept graphs. As shown in [55], this list of semantic types needs to be different depending on the medical task (diagnosis, treatment, or test). It is described in [55] that by using these concept semantic types, which filters out a large portion of concepts, the concepts extended to the query can significantly improve the retrieval precision.

In an IR task, a large of number of concepts, which are directly or indirectly related to the query, need to be examined to identify those that can improve the precision of retrieval results. In [30, 118, 117], only the concepts that are directly related to the original concepts are expanded to the original query. Not all concepts (or entities) that can increase the precision of an information retrieval system are directly related to the entities in the original query. So, in the mentioned papers, other resources to obtain related concepts are proposed. In other words, other than the concepts extracted from knowledge bases, concepts that are extracted from resources like top-ranked documents (PRF concepts) are considered as expansion concepts. The

majority of the concepts that are indirectly related to the original concepts are not useful and extracting useful and indirectly related concepts is a challenge. In [48] concepts that are directly and indirectly related to the original concepts are weighted and those with the highest weights are considered as the expansion concepts. In [48], the set of concepts that are weighted are selected from the concepts that are related to the original concepts through at most N intermediate concepts. Depending on the collection when $N > 3$ or $N > 2$ expanding new concepts only results in topic drift [48].

1.4. Concepts in Medical Knowledge-bases

In a query generated by a medical practitioner, only a portion of concepts that are required to generate accurate retrieval results are usually provided in the original query generated by the user. These concepts, which are called explicit concepts, can be identified from the original query. The rest of the concepts, which are called latent concepts, can be extracted from resources like top-ranked documents, other concepts can be obtained from external resources, like knowledge bases.

Medical domain concepts can be extracted from different resources such as SNOMED CT, UMLS, ICD-11, etc. UMLS metathesaurus is known to be the most comprehensive metathesaurus that is generated in the medical domain [80]. This metathesaurus is composed of CPT [9], ICD-10-CM [107], LOINC [68], MeSH [60], RxNorm [82], and SNOMED-CT [115]. UMLS has also another knowledge source which is called semantic network [66]. UMLS semantic network provides broad categories which are called semantic types. Relationship between semantic types are also provided by this network. The third tool is called SPECIALIST Lexicon which is a Natural Language Processing (NLP) tool [41].

Knowledge bases provide a variety of information on what are the concepts are and how they are related to each other. UMLS metathesaurus store this information by using two relational formats. The first is the Rich Release Format (RRF) and the Original Release Format (ORF). Each concept has a Concept Unique Identifier (CUI), Preferred Terms[1] (PT) Designated synonyms (SY) and so on. Each of the atoms can have one PT, SY, or other Term Types in Source (TTY). A complete list of TTYs can be found[2] in UMLS reference manual published by National Library of Medicine (NLM). For example, for concept "blood cancer", the corresponding CUI is C0376545 and one of its preferred terms is "Hematologic Neoplasms". From the provided tables in UMLS metathesaurus, the following information can be extracted:

1. Atoms of a UMLS concept: Each atom is represented by an Atom Unique Identifier (AUI). This information exists in one of the tables of UMLS called MRCONSO. For example, for concept C0376545, around 150 atoms exist in the UMLS metathesaurus. These atoms are in different languages and for each language, one of them is the preferred term [21]. As an example, some of the atoms corresponding to concept C0376545 are shown in Table 1.1. One of the signs of redundancy in UMLS methathesaurus can be observed from this table. It can be seen that atoms A18563573, A1962117 have exactly the same string. This redundancy results in an ambiguity in text annotation process. In other words, in the process of annotating concepts from a free text, there are more than one candidate that

[1] https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_004.html
[2] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html

can for the annotate concept [88]. Therefore, mapping of strings to their AUIs requires a word sense disambiguation process [81, 42].

Table 1.1. Shorted list of 150 Atoms' string, Atomic Unique Identifier (AUI), Root Source Abbreviation (RSAB), Term Type in Source (TTY) and their UMLS object's Code for the concept C0376545 with preferred term \Hematologic Neoplasms". Resources that are used to get concepts in the following table are Collaborative Consumer Health Vocabulary (CHR) [97], Computer Retrieval of Information on Scientific Projects (CSP) [10], Human Phenotype Ontology (HPO) [10], International Classification of Primary Care 2 (ICPC2P) [18], MDR and MDRCZE [23].

string	AUI	RSAB	TTY	Code
Blood cancer	A18563573	CHV	PT	0000031196
blood cancer	A18600614	CHV	SY	0000031196
bone carcinoma marrow	A18605945	CHV	SY	0000049723
bone marrow carcinoma	A18661683	CHV	PT	0000049723
cancer blood	A18563574	CHV	SY	0000031196
carcinoma bone marrow	A18661684	CHV	SY	0000049723
hematologic cancer	A18581974	CHV	SY	0000031196
hematologic malignancies	A18582166	CHV	SY	0000031810
hematologic malignancy	A18619355	CHV	PT	0000031810
hematologic neoplasms	A18638055	CHV	SY	0000031810
hematological malignancies	A18582167	CHV	SY	0000031810
hematological malignancy	A18675129	CHV	SY	0000031810
malignancies hematologic	A18563761	CHV	SY	0000031810
…	…	…	…	…

2. Definitions of a UMLS concept: All of the definitions from different knowledge sources for a UMLS concept is provided in a UMLS table called MRDEF. For example, the definition provided for the concept C0376545 is: "*Neoplasms located in the blood and blood-forming tissue (the bone marrow and lymphatic tissue). The most common forms are the various types of LEUKEMIA, of LYMPHOMA, and of the progressive, life-threatening forms of the MYELODYSPLASTIC SYNDROMES*." This definition is provided by MESH, which is one of the knowledge sources of UMLS.

Table 1.2. List of concepts related to concept C0376545 with preferred term “Hematologic Neoplasms”. In this table, REL denotes the concept relationship and RSAB denotes Root Source Abbreviation. RB is an abbreviation for broader relationship, RN is abbreviation for narrower relationship and RO is an abbreviation for having relationship other than synonymous, narrower, or broader. MTH stands for MeSH knowledge source.

CUI	String	REL	RSAB
C0376544	Hematopoietic Neoplasms	RB	MTH
C3890429	Liquid Tumor	RO	MTH
C0348393	Malignant tumor of lymphoid hemopoietic and related tissue	RN	MTH

3. **Relationship and inverse relationships of a UMLS concept:** For the concept C0376545, the related concepts can be found in Table 1.2. It will be explained later that the concept relationships table is one of the most popular resources for query expansion. Query expansion process needs to obtain concepts that are related to the query concepts and also be useful for the query expansion [7]. In other words, it is not guaranteed that if a query expanded with its related concepts, the retrieval quality will improve. It is mainly because of redundancy in UMLS metathesaurus and not being UMLS designed for query expansion purposes. Some of the concepts do not have appropriate concept strings for the task of query expansion. For example, as can be seen from this table, the concept C0348393 with string “*Malignant tumor of lymphoid hemopoietic and related tissue*” has terms like “*and related*" that is not useful if added to the original query. Some strings have strings that has explanations about the concepts and are not good representative of their corresponding concepts [67]. On the other hand, the comprehensiveness of knowledge bases comes at the expense of their large dimensionality, redundancy. A very large number of concepts can be directly or indirectly related to a query, but only a small fraction of them are effective for query expansion. Recent decade has witnessed the

emergence of a large number of general purpose and domain-specific on-line knowledge bases.

Table 1.3 List of semantic types considered in [55] in the query expansion process. The concepts related to the query with semantic types mentioned in this table are considered as expansion concepts in [55]. In this table, STY is an abbreviation for semantic type and TUI for Type Unique Identifier.

STY	TUI	Semantic type
blor	T029	Body Location or Region
bpoc	T023	Body Part, Organ, or Organ Component
clnd	T200	Clinical Drug
diap	T060	Diagnostic Procedure
diap	T060	Disease or Syndrome
fndg	T033	Finding
hlca	T058	Health Care Activity
inpo	T037	Injury or Poisoning
inpr	T170	Intellectual Product
medd	T074	Medical Device
mobd	T048	Mental or Behavioral Dysfunction
neop	T191	Neoplastic Process
patf	T046	Pathologic Function
phsu	T121	Pharmacologic Substance
sosy	T184	Sign or Symptom
top	T061	Therapeutic or Preventive Procedure

4. Semantic types of the concepts and relationship between them: There are over 100 semantic types and their relationships in the UMLS semantic network [66]. Not all of the concepts that related to the query concepts are useful for the query expansion task. One of the factors that can be considered to narrow down the options to more useful ones is to use semantic types of concepts [110, 55, 2, 45]. Different list of semantic types is suggested, such as the one shown in Table 1.3, to filter out UMLS concept for medical query expansion [55].

1.5. Overview of our Query Expansion Methods

We aim at improving the medical IR models that utilize knowledge bases for CDS and PM. The first method in this dissertation represents medical concepts extracted from verbose medical queries and knowledge bases. This method accounts for the differences in the importance of different knowledge bases in representing the medical query concepts. Next, we describe our method to compute the weights of a medical IR model with the goal of optimizing the retrieval performance. Then, we present our method to extract concepts from a concept graph with the objective of minimizing the number of evaluated concepts by keeping the retrieval performance above a certain threshold. The last method that we describe in this dissertation utilizes a Bayesian approach to utilize knowledge bases and perform query expansion in an IR task for PM.

Utilization and Impact of this research in clinical practice: Our query expansion methods together with an IR system can help clinicians in their access to a collection of medical articles (such as PubMed) or a collection of Electronic Medical Records (EMRs) given a patient case in the form of the query. These methods can fill the vocabulary gap between an underspecified medical query and its relevant documents. They can be utilized in medical-domain search engines such as PubMed to improve the quality of the retrieval systems. The impact of these method can be significant under the following four scenarios:

1. The medical queries are verbose.
2. An effective query expansion requires a diversity of knowledge bases.
3. Good expansion concepts (Concepts that can improve the retrieval performance) are indirectly related to the query concepts through intermediate concepts.
4. An effective query expansion requires a prior knowledge about relatedness of expansion concepts to the query concepts.

CHAPTER 2 RELATED WORK

Depending on the type of concepts used for query expansion, general-purpose and domain-specific retrieval methods can be categorized into the ones that are based on statistical concepts (i.e., determined based on term popularity and co-occurrence in a given collection) [15, 17, 69, 70, 104], the ones that are based on semantic concepts (i.e., that are extracted from a knowledge repository) [55, 101, 103, 128], and those that combine semantic and statistical concepts [27, 94, 116, 45]. Below we provide an overview of the previously proposed methods in each of these three categories.

2.1. Retrieval methods using statistical concepts.

In the simplest case, these retrieval models utilize only unigrams from the top retrieved documents for query expansion [104]. More recent retrieval methods utilizing statistical concepts are based on the Markov Random Field (MRF) framework introduced by Metzler and Croft [69]. It assigns the same importance weight to all matching statistical query concepts of the same type (unigrams and sequential bigrams), when the retrieval score of a document is calculated. Latent Concept Expansion (LCE) extends MRF by also using unigrams from the PRF documents as latent concepts for query expansion. The requirement of having fixed weights for unigrams and bigram concepts in the MRF-based retrieval model was relaxed by the Weighted Sequential Dependence (WSD) model [16], which estimates the importance of each concept individually. A similar relaxation of LCE weights was implemented in the Parameterized Query Expansion (PQE) [17] model. Overall, query representation methods based on statistical concepts typically consider unigrams and bigrams in the query and/or unigrams in PRF documents.

2.2. Retrieval methods using semantic concepts.

Semantic concepts for query expansion are typically extracted from domain-specific, such as the Unified Medical Language System (UMLS) [55], Medical Subject Headings (MeSH) [61] and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [46], or general-purpose knowledge repositories, such as Wikipedia [101, 119]. The utility of this type of concepts has been studied for a variety of medical IR tasks including medical literature retrieval [98, 129]. UMLS concepts are typically extracted from queries and top-retrieved documents using MetaMap [6, 35, 55, 100, 101, 111, 112].

Soldaini et al. [101] proposed two methods for medical literature retrieval that use Wikipedia-based heuristics to filter out non-medical concepts from the original query and top retrieved documents. The first method (referred to as HT in [101] and Wiki-Orig in this dissertation) is a query reduction method, which retains only those bigram concepts in the original query that are determined to be health-related according to a heuristic. On the other hand, the second method (referred to as HT-PRF in [101] and Wiki-TD in this dissertation) expands the original query with a number of health-related concepts that are extracted from the top-retrieved documents and filtered out using the same heuristic.

Accounting for semantic types of concepts[3] can also significantly improve the accuracy of query expansion, as they can be used to filter out the candidate expansion concepts. The method proposed in [55] (referred to as UMLS-TD in this work), expands medical queries only with the UMLS concepts extracted from the top retrieved documents that have pre-selected semantic types. A semantic type is pre-selected if the concepts of this type improve the accuracy of retrieval results when added to the queries in the training set. For example, the semantic type

[3] http://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

"signs and symptoms" is pre-selected for a query about the diagnosis of a disease. [111] proposed another approach to using semantic types in the query, in which the semantic types of concepts are used to weight the concepts (concepts that are more likely to be effective, get higher weight).

Retrieval models using both semantic and statistical concepts. The benefit of integrating semantic and statistical concepts was shown in [27, 45, 94, 116]. The methods in [27, 94, 116] focused only on explicit concepts (query unigrams and bigrams along with UMLS concepts extracted from the query using MetaMap). A medical IR system that integrates a graph-based representation of the corpus, structured knowledge sources and a retrieval model combining statistical IR methods with an inference mechanism implemented as graph traversal has been proposed in [45].

2.3. Concept graphs for query expansion

Concept graphs are widely used in domain-specific [12] and general-purpose [30, 48] information retrieval (IR) systems. They provide structured knowledge that is necessary to fill in the gap between the information provided by a user in the form of a query and the information required by a retrieval system in order to return complete and accurate results. Concept graphs can be constructed from a document collection as in [11, 47, 48]; semantic network, such as ConceptNet [11, 48]; or from an entity-centric knowledge graph, such as DBpedia [11] or Freebase [11]. Since there can be a very large number of concepts in a concept graph that are related to a query, traditional methods for concept selection from the top retrieved documents, such as the one proposed in [25] and [119], that exhaustively evaluate all candidate concepts can be quite inefficient.

To tackle the difficulty of examining a large number of concepts, simple approaches [55, 101] utilizing external information to prune useless expansion concepts have been previously proposed for domain-specific IR. Experimental evaluation of these methods have shown that it is possible to achieve a significant improvement in retrieval accuracy by pruning the candidate concepts with certain properties, such as semantic types. In particular, a medical IR system proposed in [101] discards candidate expansion concepts from the top retrieved documents that are determined to be unrelated to healthcare based on a simple Wikipedia-based heuristic. The method proposed in [55] does not consider the candidate concepts from the Unified Medical Language System concept graph, the semantic type of which does not belong to a pre-determined list of semantic types known to be effective for specific medical tasks associated with medical record search queries. Since general-purpose retrieval systems operate with a larger and more diverse set of concept and query types than domain-specific ones, they cannot effectively prune candidate expansion concepts based on simple heuristics.

Query expansion methods utilizing general-purpose entity-centric knowledge graphs, such as DBpedia and Freebase, have been extensively investigated in recent years [30, 117, 118, 119]. These methods require annotations of the queries (and, in some cases, also of the documents) with links to Freebase entities, which makes them ineffective for the queries that are ambiguous, broad or do not contain proper nouns designating named entities that can be linked to a knowledge graph.

Kotov and Zhai [48] studied the retrieval effectiveness of expansion concepts from ConceptNet that are related to the query concepts thorough one or several intermediate concepts. In particular, their method first sorts all ConceptNet concepts, which are related to the

query concepts through at most 2 intermediate concepts, according to predicted average precision AP) of retrieval results after adding each concept and uses the top 100 concepts with the highest predicted AP to create a query expansion language model. They found out that, although the majority of the concepts in the second and third concept layers do not improve the accuracy of retrieval results, there are several highly effective concepts in these layers. However, finding them requires evaluation of a large number of concepts.

Sequential analysis (and active learning, its closely related area) have been adopted by many methods to deal with very large datasets. These methods aim to minimize the cost (or time) spent on obtaining reasonably accurate results. In IR, these methods have been applied to minimize (or reduce) the relevance feedback effort (i.e., the number of relevance judgments of retrieved documents), while maintaining an acceptable level of retrieval accuracy [54, 108, 125]. Diaz [31] proposed a method that sequentially selects query expansion terms from the top retrieved documents and achieves a significant improvement over standard pseudo-relevance feedback (PRF) approaches.

CHAPTER 3 Representing Concepts in Medical Information Retrieval

3.1. Introduction

In this chapter, we present a Markov Random Fields-based retrieval model and an optimization method for jointly weighting statistical and semantic unigram, bigram and multi-phrase concepts from the query and PRF documents as well as three specific instantiations of this model that we used to obtain the runs submitted for each task in TREC 2015 Clinical Decision Support (CDS) track. These instantiations consider different types of concepts and use different parts of topics as queries.

Previously proposed approaches to identify and weight query concepts are either based only on semantics [55, 101, 44, 103, 128] or are purely statistical [104, 69, 15, 17, 49, 70, 127]. Each of these two types of approaches are able to identify only certain types of concepts. For example, [55] identifies and utilizes only the concepts from the Unified Medical Language System (UMLS) that are extracted using the MetaMap tool [5] from PRF documents. Single-word and multi-word statistical concepts from the query and single-word concepts from PRF documents have been shown to be effective for ad-hoc retrieval in [70, 17]. A bag-of-words retrieval model utilizing medical concepts from PRF documents for query expansion was proposed in [101]. Choi et al. [27] proposed a method to represent multi-word UMLS concepts using sequential dependencies between their words.

In this chapter, we present a Markov Random Fields-based retrieval model and an optimization method for jointly weighting statistical and semantic unigram, bigram and multi-phrase concepts from the query and PRF documents as well as three specific instantiations of this model that we used to obtain the runs submitted for each task in TREC 2015 Clinical Decision

Support (CDS) track. These instantiations consider different types of concepts and use different parts of topics as queries.

3.2. Method

In this section, we provide the details of the six runs that were submitted to TREC 2015 CDS track. Three of these runs were submitted for Task A and three others were submitted for Task B of this track. The runs submitted for Task B consider the diagnosis section provided for some of the topics in this task. These diagnosis sections are considered as n-gram concepts and added with the optimized weights to the expanded queries. As mentioned in [95], considering all of the runs in TREC 2014 CDS track, a very small difference in retrieval performance is observed when the query types (i.e., “Diagnosis”, “Test”, and “Treatment”) are taken into account. Therefore, query types are not taken into account in this work.

In this work, we assume that the concepts representing the information need underlying the query exist both in the query itself as well as in other concept sources, such as PRF documents. We also assume the existence of sequential dependencies between the adjacent terms of multi-word concepts, which can be accounted for in retrieval by using the Markov Random Field (MRF) model [69]. In particular, our retrieval model builds upon the Markov Random Field-based Parameterized Query Expansion (PQE) framework [17], which assumes that the information need underlying a multi-term query can be categorized using three query concept types (unigrams, ordered bigrams, and unordered bigrams), each of which is associated with its own matching function. We extend this framework by considering more fine-grained concept types, depending on whether the concepts of the above three types occur in the query itself (including the multi-word UMLS concepts) or in the PRF documents, and thus providing a

more flexible concept matching strategy. Specifically, in our retrieval model, contribution of a query concept c to the retrieval score of document D, in which it occurs, is determined as:

$$sc(c, D) = \sum_{T \in \mathbf{T}} \lambda_T f_T(c, D) \quad (3.1)$$

where $\mathbf{T}$ is a set of all concept types, to which concept c belongs (a query concept can belong to several concept types; for example, if it occurs in both the query and the PRF documents) and λ_T is the relative importance weight of the concepts of type T (all concepts of the same type are assigned the same weight). The final retrieval score of document D given a query is determined as a linear combination of contributions of all query concepts occurring in D:

$$sc(Q, D) = \sum_{c \in \mathbf{C}} \mathrm{I}_c \; sc(c, D) = \sum_{c \in \mathbf{C}} \mathrm{I}_c \; \sum_{T \in \mathbf{T}} \lambda_T f_T \; (c, D) \quad (3.2)$$

where $\mathbf{C}$ is the set of all explicit and latent query concepts, I_c is an indicator function that determines whether concept c is considered (i.e., it takes the value of 1) or not (i.e., it takes the value of 0). In other words, concept types are weighted, but individual query concepts can be used or discarded. The query set and relevance judgments from TREC 2014 CDS track were used to optimize concept importance weights and other parameters of the models.

Concept types

The methods that were used to obtain the 6 runs submitted to the CDS track are summarized in Table 3.1. Besides the type (manual or automatic) and part of the topic that they used as a query, these methods are different by the query concept types they consider.

Overall, the submitted runs utilize 4 concept sources: the query itself, PRF documents, Unified Medical Language System (UMLS) concepts extracted from the query and Google search results. Query terms, PRF documents and UMLS concepts are used by the automatic methods. For manual methods, (i.e., wsuirdma, wsuirsmb and wsuirdmb), we manually extracted a number

Table 3.1: Summary of retrieval runs submitted to TREC 2015 CDS track.

Method	Query	Method	Task
wsuirsaa	summary	automatic	A
wsuirdaa	description	automatic	A
wsuirdma	description	manual	A
wsuirsab	summary	automatic	B
wsuirsmb	summary	manual	B
wsuirdmb	description	manual	B

of concepts from Google search results and added them to the expanded query, in addition to the concepts from the other 3 sources. Concept types from different sources that were used by different retrieval runs are summarized in Table 3.2.

All unigram concepts extracted from the original query are retained in the transformed query. Since the top retrieved documents may or may not be relevant to the original query, only a small number of unigram concepts with the highest weight in the relevance model [51] were added to the original query. The optimal number of these concepts was determined using the training data. UMLS concepts (which can consist of more than two terms) were extracted from the query using the MetaMap tool [5]. Multi-word UMLS query concepts were broken down into sequential bigrams. For example, a multi-word concept "Iron Deficiency Anemia" was represented using the Indri query language as follows:

```
1.00 #weight(
0.40 #combine( Iron Deficiency Anemia )
  0.35 #combine( #od4( Iron Deficiency )
            #od4( Deficiency Anemia ) )
  0.45 #combine( #uw17( Iron Deficiency )
            #uw17( Deficiency Anemia ) )
)
```

where 0.40, 0.35 and 0.45 are the weights of the corresponding concept types. The window sizes for ordered and unordered bigrams (i.e., 4 and 17, respectively) were determined to optimal

based on the training data. It is notable that it is not necessary to normalize the mentioned weights in Indri query language to be sum-to-one as this normalization is done automatically by Indri.

Since it was shown in previous work [93] that UMLS concepts may or may not improve the performance of the medical information retrieval, only the concepts that belong to the following semantic types (https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt) are included in the expanded query:

- Clinical Drug
- Disease or Syndrome
- Injury or Poisoning
- Sign or Symptom
- Therapeutic or Preventive Procedure

This list was obtained from an initial list of 16 semantic types in [55] through backward elimination process [73]. Unlike [55], in which the list of considered concept types is different for each query type (i.e., “Symptom”, “Diagnostic test”, “Diagnosis” and “Treatment” queries), we considered the same semantic types for “Diagnosis”, “Test” and “Treatment” queries.

A number of concepts that were added to the original queries in manual runs were selected from the top 10 Google search results. This selection process is done manually from the content of the documents retrieved by the Google web search engine in response to the summary or description fields of TREC CDS topics used as queries. In the case of narrative queries, the queries were modified slightly to increase the recall in Google search. Only healthcare-related concepts that are relevant to the information need of queries were added to them. The number

of concepts that are extracted from Google search results and added to the transformed query depends on the relevance of documents in search results. Two factors that are considered in manually selecting the concepts from Google search results are:

1. relatedness of these concepts to the medical domain (e.g., “Kawasaki disease” is a highly related concept),

2. popularity of these concepts in medical domain (e.g., “health care” is too popular in the medical domain).

In other words, the desired concepts for query expansion in this case are the ones that are highly related to the medical domain, but not too popular.

Each concept type has different weight, as determined by its level of importance in the query. Intuitively, unigram query concepts are typically more important than unigram concepts from PRF documents. Therefore, choosing appropriate concept weights in (4.1) is a very important step in query transformation. We used Coordinate Ascent [71] to estimate those weights on the training data. In this optimization method, the weights are optimized one after another until convergence.

3.3. Experiments

All the runs reported in this work were obtained using Indri 5.72 [105] IR toolkit. A two-stage document language model smoothing method proposed in [124] was used in conjunction with all retrieval models. The accuracy of 6 submitted runs in terms of Inferred Average Precision (infAP), Inferred Normalized Discounted Cumulated Gain (infNDCG), R-precision (R-prec), Precision at 10 (P@10), and Mean Average Precision (MAP) is summarized in Table 3.3.

Experimental results in Table 3.3 lead to several conclusions. First, we observe that wsuirdma, which is a manual method using unigrams from topic descriptions, PRF documents and Google search results, as well as ordered and unordered bigrams from UMLS concepts in topic descriptions and Google search results, has the highest performance in terms of all metrics for Task A of the CDS track. Second, we observe that wsuirdaa, which is an automatic method

Table 3.2 Concept types utilized by submitted retrieval runs.

Concept Types	wsuirsaa	wsuirdaa	wsuirdma	wsuirsab	wsuirsmb	wsuirdmb
unigrams in topic summary	●			●	●	
ordered bigrams in UMLS concepts in topic summary	●			●	●	
unordered bigrams in UMLS concepts in topic summary	●			●	●	
unigrams in topic description		●	●			●
ordered bigrams in UMLS concepts in topic description		●	●			●
unordered bigrams in UMLS concepts in topic description		●	●			●
unigrams in PRF documents	●	●	●	●	●	●
unigrams in Google search results			●		●	●
ordered bigrams in Google search results			●		●	●
unordered bigrams in Google search results			●		●	●
unigrams in diagnosis field				●	●	●
ordered bigrams in diagnosis field				●	●	●
unordered bigrams in diagnosis field				●	●	●

using topic descriptions as queries, outperforms wsuirsaa, which is another automatic method using topic summaries as queries. Similarly, for Task B, wsuirdmb, which is a manual method using topic descriptions as queries has significantly better retrieval accuracy than wsuirsmb, which is using topic summaries as queries. Third, we observe that incorporating information about diagnosis of the disease, which is provided in Task B, generally increases the retrieval accuracy of our models, particularly the manual ones.

Table 3.3 Summary of performance for all submitted runs.

Methods	infAP	infNDCG	R-prec	P@10	MAP
wsuirsaa	0.0777	0.2928	0.2329	0.4633	0.1851
wsuirdaa	0.0842	0.2939	0.2306	0.4667	0.1864
wsuirdma	0.0880	0.3109	0.2493	0.4733	0.1968
wsuirsab	0.0875	0.3246	0.2656	0.5067	0.2180
wsuirsmb	0.0856	0.3208	0.2608	0.5033	0.2116
wsuirdmb	0.1014	0.3690	0.2843	0.5200	0.2331

Topic-level differences in terms of infNDCG between our best automatic and manual runs and the median performance of the corresponding runs submitted to the CDS track by all other teams for Task A and Task B are illustrated in Figures 3.1 and 3.2, respectively. For Task A, our best automatic and manual runs have greater infNDCG than the median for 22 out of 30 topics (73.33%). For Task B, our best automatic run has greater infNDCG than the median for 70% of the topics, our best manual run for this task has greater infNDCG than the median for 86.67% of the topics and is slightly worse than the median for only 4 topics.

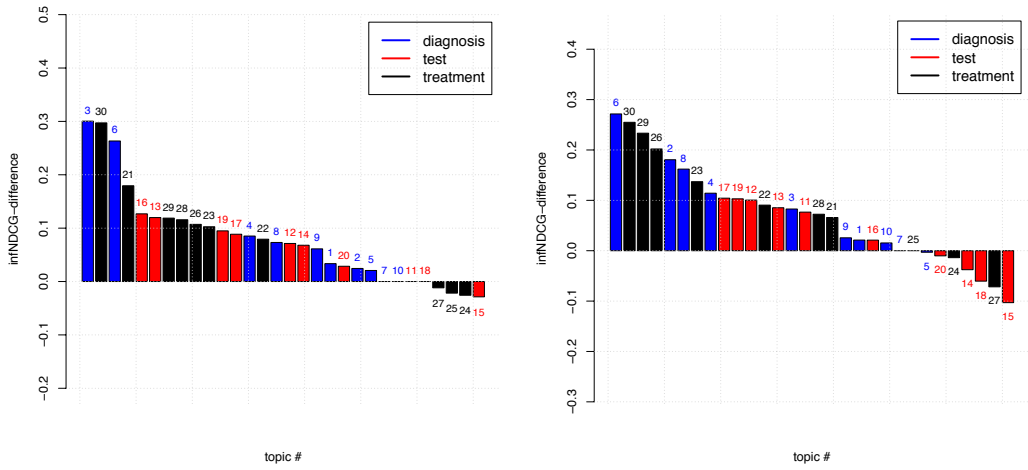


Figure 3.1: Topic-level differences in terms of infNDCG between the proposed manual and automatic methods and the median for all TREC 2015 CDS track runs for Task A.

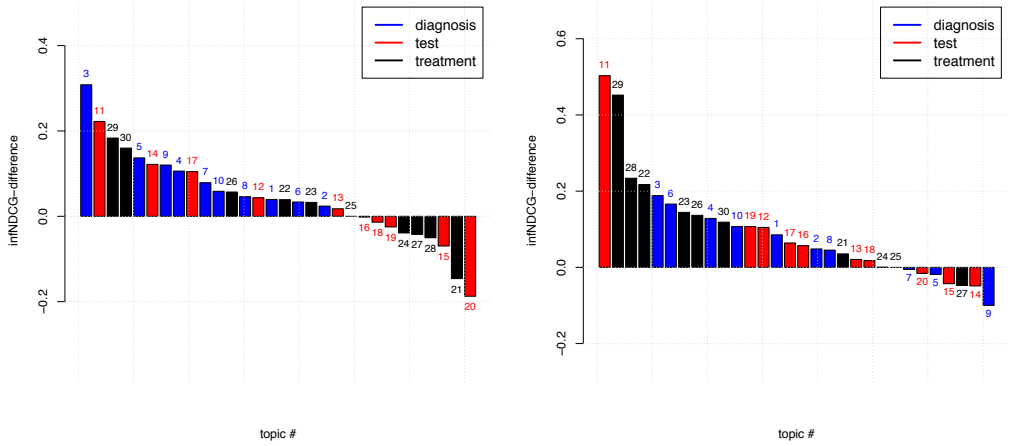


Figure 3.2: Topic-level differences in terms of infNDCG between the proposed manual and automatic methods and the median for all TREC 2015 CDS track runs for Task B.

CHAPTER 4 AN OPTIMIZATION TECHNIQUE TO WEIGHT EXPANSION CONCEPTS FROM KNOWLEDGE BASES

4.1. Introduction

Given descriptive summary of a medical case as a query, the goal of information retrieval systems for clinical decision support (CDS) is to return articles from a collection of medical literature that are relevant to the query and can assist a clinician in making decisions regarding the case, such as prescribing a medication, procedure or treatment. A fundamental challenge faced by those systems is that although CDS queries are typically verbose and may consist of several sentences (e.g. "*33-year-old male presents with severe abdominal pain one week after a bike accident, in which he sustained abdominal trauma. He is hypotensive and tachycardic, and imaging reveals a ruptured spleen and intraperitoneal hemorrhage*"), only a small subset of query terms (henceforth referred to as explicit concepts) correspond to the key query concepts, such as "*bike accident*", "*abdominal trauma*", "*tachycardia*", "*splenic rupture*", "*intraperitoneal hemorrhage*", which represent the information need behind this query, while many other important concepts that are relevant to this information need (e.g. "*spontaneous spleen rupture*", "*splenic trauma*", etc.) are not directly mentioned in the query (henceforth referred to as latent concepts). Providing complete and accurate retrieval results for CDS queries requires both correct identification of the key explicit concepts and addition of important latent concepts to the query, as well as precise weighting of explicit and latent concepts in the modified query.

In this chapter, we describe our method to represent verbose clinical decision support queries using unigram, bigram and multi-term concepts from the query itself, as well as from the PRF documents and external knowledge bases (such as the Unified Medical Language System).

Our method is based on linear feature-based learning-to-rank retrieval framework [71], in which the relative importance weight is determined for each matching query concept individually as a linear combination of features. We also propose a set of features for each concept type, which is determined based on whether a concept is a unigram, bigram or multi-term phrase and whether it occurs in the query itself or is extracted from a top retrieved document or a knowledge base.

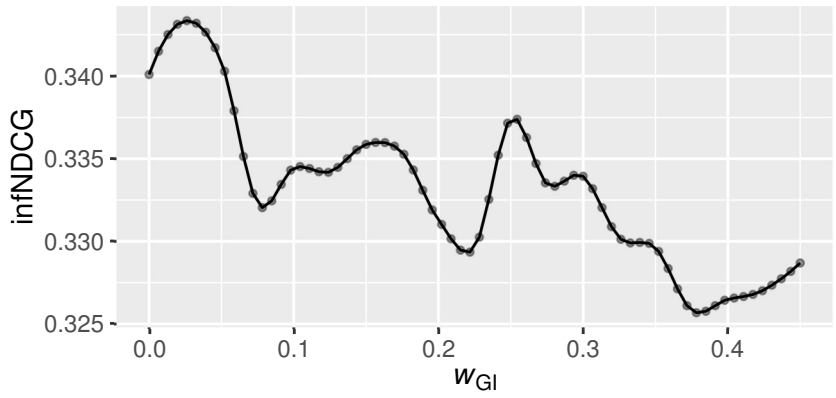


Figure 4.1: The values of objective function corresponding to infNDCG retrieval metric by varying the weight of one of the features (GI presented in Section 6.2), which determines the importance of concept matches of certain type.

Since the parameter spaces of linear feature-based retrieval models can be reduced to a multinomial manifold, their parameters can be estimated by direct maximization of the target rank-based retrieval metric (e.g. NDCG) over this manifold using derivative-free unconstrained multi-dimensional optimization methods, such as coordinate ascent [72] or hill-climbing [76]. These methods are based on the Powell's method, which divides a complex multi-dimensional optimization problem into several simple one-dimensional ones. After that, it iteratively optimizes a multivariate objective function by optimizing each parameter individually, while holding all other parameters fixed. Since line search is a local optimization method, the efficiency and accuracy of both the coordinate ascent and hill-climbing rely on the assumption of

smoothness and convexity of objective function when a free parameter is optimized, which is often violated in practice. Figure 4.1, which shows the behavior of the target retrieval metric by varying the value of a parameter that corresponds to the weight of a feature, illustrates this case. It can be seen that the objective function shown in this figure has several local maxima.

The optimization method for learning the weights of concept importance features in feature-based retrieval models proposed in this paper leverages the Graduated Non-Convexity (GNC) (or continuation) optimization method [20] to address the issue of non-smooth and non-convex objective functions, when individual parameters are optimized using the Powell's method. GNC is a derivative-free method specifically designed for global optimization of non-smooth and non-convex objective functions. Graduated Non-Convexity (GNC) is an iterative method, which applies different degrees of smoothing to the original objective function to generate smoother and more convex objective functions, which have their global maximum close to the one of the original objective functions. The method starts by applying the highest degree of smoothing and then gradually decreases the rate of smoothing at each subsequent iteration using the result obtained at the previous iteration as the starting point for the next iteration until the global maximum for the original non-smoothed objective function is found. Although the quality of the solution attained by this approach heavily depends on the choice of the smoothing method, it was recently shown that Gaussian smoothing of a non-convex function is optimal in a sense that it evolves any function into its convex envelope [75].

The key difference of the proposed method from existing methods for medical literature and ad hoc document retrieval is that it uses both statistical and semantic concepts extracted from diverse sources (query itself, knowledge bases and top retrieved documents) for query

representation. The proposed method also leverages an efficient optimization technique to learn the relative importance weight of different types of query concepts on the same scale.

4.2. Method

In this section, we present the details of the proposed query reformulation method, a set of features used with it and a method to optimize the weights of those features with respect to the target retrieval metric. The proposed query reformulation method combines explicit and latent query concepts from diverse sources and determines the weight of each individual concept as a linear combination of features, which depend on a concept type. The type of a query concept is determined by its source and whether the concept is represented by a unigram, bigram or multi-word phase. The set of concept sources considered in our method includes the query itself, top retrieved documents for the original query, and external knowledge repositories.

Retrieval Model

To account for term dependencies, the proposed method adopts a Markov Random Field (MRF) retrieval framework [69], in which the retrieval score of a document is determined as a weighted linear combination of the matching scores of different concept types in a given query. In particular, our method extends the parametrized concept retrieval model in [17], according to which the retrieval score of document D with respect to query Q is calculated as:

$$sc(Q,D) = \sum_{T \in T_Q} \sum_{c \in C_T} \lambda_T(c) f_T(c,D) \tag{4.1}$$

where C_T is a set of concepts belonging to concept type T, and $\lambda_T(c)$ is defined as the importance weight of concept c, which depends on its type. In the above equation, $f_T(c,D)$ is the matching score function of concept c in document D, which is defined as:

$$f_T(c,D) = \log\left((1-\lambda)\frac{n(c,D)+\mu\frac{n(c,Col)}{|col|}}{|D|+\mu} + \lambda\frac{n(c,Col)}{|Col|}\right) \tag{4.2}$$

where $n(c, D)$ $(n(c, Col))$ and $|D|$ $(|Col|)$ are the counts of concept c in document D (entire collection) and the size of document D (entire collection), respectively. The above matching function utilizes a two-stage smoothing method from [124], where λ and μ are Jelinek-Mercer and Dirichlet smoothing coefficients, respectively. Since only unigrams as well as ordered and unordered bigrams are considered in the MRF retrieval framework, concepts that are represented by multi-word phrases are broken down into unigrams and sequential bigrams. The set of concept types considered for a query Q is designated by T_Q and is shown in Table 4.2. This table also provides information about the concept extraction methods and a set of features corresponding to each concept type, which will be explained in detail below.

The importance weight of concept c is parameterized using a set of importance features $\Phi_T(c)$. Each concept type T is associated with its own set of importance features, summarized in Table 5.4. Thus, the weight of concept c with type T is determined as a weighted linear combination of importance features:

$$\lambda_T(c) = \sum_{n=1}^{N} w_\phi^n \phi_n \,, \tag{4.3}$$

where $\{\phi_1, \dots, \phi_N\}$ is a set of features for concepts with type T (i.e., $\Phi_T(c) = \{\phi_1, \dots, \phi_N\}$), and w_ϕ^n is the importance weight of the n-th feature (i.e., ϕ_n). The intuition behind this concept weighting scheme is that different concept types have different importance and should be weighted accordingly. Intuitively, knowledge-based concepts (such as the UMLS concepts) that are linked from the concepts in the original query should have a different importance weight than the concepts that are extracted from the top retrieved documents. Similarly, bigrams corresponding to UMLS concepts identified in the original query should be weighted differently than other bigrams in the original query. On the other hand, features determining the importance

of a concept from a graph structured knowledge repository (e.g. UMLS), like the degree of the node corresponding to this concept, are different from the features that determine the importance of a unigram concept in top retrieved documents.

Optimization Method

Learning the feature weights that maximize the target retrieval metric on a training data can be considered as a multivariate optimization problem and is typically addressed by decomposing it into a set of one-dimensional optimization problems. Instead of performing a line search along every single dimension in optimizing a set of feature weights with respect to the target retrieval metric, we propose to use graduated optimization [20], an efficient global optimization technique.

Graduated optimization

Graduated optimization is an iterative optimization method that gradually finds the global optimum of a given objective function by finding the optima for a series of simplified objective functions. Each of these simplified objective functions is obtained from the original objective function by applying different degree of smoothing to make the original function more convex. It starts from the solution to the most simplified optimization problem (i.e., when the maximum degree of smoothing is applied to the original objective function) and considers this solution as the starting point for the second less simplified problem (i.e., less smoothed original objective function). This process continues until the global optimum for the original objective function is found. This procedure is based on the assumption that the global optimum of a given objective function at the current iteration is close enough to its global optimum at the next iteration. Therefore, at the next iteration, the region of the parameter space that is far enough from the

optimum point at the current iteration is ignored. As a result, a smaller region that is close to the optimum point at the current iteration is searched for the optimal parameter setting at the next iteration.

Smoothing method

In case of a univariate optimization problem with a single parameter w_ϕ, the smoothed objective function, $\tilde{E}(w_\phi)$, can be obtained by taking sample values from $E(w_\phi)$, the original objective function. To compute $\tilde{E}(w_\phi)$ at a specific region around the starting point $w_{\phi,0}$, samples are taken from $\tilde{E}(w_\phi)$ for the following values of w_ϕ:

$$\boldsymbol{w}_{s,\phi} = [w_{\phi,-M}, \dots, w_{\phi,0}, w_{\phi,M}] \tag{4.4}$$

where

$$w_{\phi,m} = w_{\phi,0} + m\Delta w_\phi, \qquad m \in [-M, \dots, M] \tag{4.5}$$

and Δw_ϕ is the sampling interval.

When a polynomial of degree K is used for the smoothed objective function at point $w_{\phi,m}$:

$$\tilde{E}\left(w_{\phi,m}\right) = \sum_{k=0}^{K} a_k m^k, \qquad m \in [-M, \dots, M] \tag{4.6}$$

The weight $a_k$is determined so that the following Mean Square Error (MSE) is minimized:

$$\varepsilon_\phi = \frac{1}{2M+1} \sum_{m=-M}^{M} \left(\tilde{E}(w_{\phi,m}) - E(w_{\phi,m})\right)^2 \tag{4.7}$$

As shown in [92], optimal $a = [a_1, \dots, a_M]$ is found as:

$$a = (\boldsymbol{J}\text{^}T\boldsymbol{J})^{-1}\boldsymbol{J}^T\boldsymbol{w}_{s,\phi}, \tag{4.8}$$

where $\boldsymbol{J}$ is a Jacobian of the vector $[\tilde{E}\left(w_{\phi,-M}\right), \dots, \tilde{E}(w_{\phi,M})]$, and its (m,k)-th element is obtained as:

$$[\boldsymbol{J}]_{m,k} = (m - M)^k, \qquad m \in [0, 2M],\ k \in [0, K]. \tag{4.9}$$

where $M, \Delta w_\phi$ and K control the smoothing rate of the objective function.

Figure 4.2 illustrates three iterations of the smoothing procedure to find the optimal weight for one of the features (w_{GI}). Points in Figure 4.2 indicate the samples taken from the objective function at each iteration, while the solid lines indicate the smoothed curves (i.e., estimated polynomials). The maximum of the smoothed curve is found and used as the starting point for the next iteration. At each subsequent iteration, the degree of smoothing is reduced by lowering Δw from 2.5×10^{-2} to 2.5×10^{-3} and then to 2.5×10^{-4}, while increasing K from 4 to 5 and 6, while keeping M constant ($M = 18$). As follows from Figure 4.2, the smoothing standard deviation (σ) is decreasing at each iteration of the optimization process, which indicates less smoothing and hence closer representation to the original objective function.

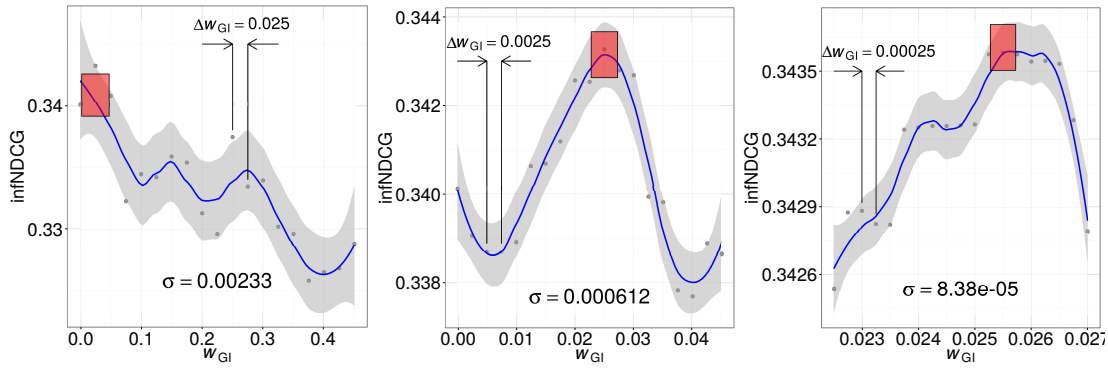


Figure 4.2: Application of graduated optimization to estimate the weight of the feature GI using TREC 2014 CDS track queries as the training set. Red boxes indicate the range of w_{GI} considered at the next iteration. σ is defined as the smoothing standard deviation.

Multi-variate optimization

The multivariate optimization method to train the weights of all features with respect to the target retrieval metric is summarized in Algorithm 1. We denote the vector of feature weights by $w_\phi = \left[w_\phi^n\right]_{n=1}^N$. As mentioned earlier, the weight w_ϕ^n is estimated by using $n - 1$ previously

estimated weights at iteration j (i.e., $\widehat{w}_{\phi}^{1}, \dots, \widehat{w}_{\phi}^{n-1}$) and the $N-n$ estimated weights at the iteration $j-1$ (i.e $\widehat{w}_{\phi}^{n+1}, \dots, \widehat{w}_{\phi}^{N}$). Therefore, the univariate objective function to estimate the weight w_{ϕ}^{n} can be written as:

$$E^{n,j}\left(w_{\phi}^{n}\right) = E\left(\left[\widehat{w}_{\phi}^{1}, \dots, \widehat{w}_{\phi}^{n-1}, \widehat{w}_{\phi}^{n}, \widehat{w}_{\phi}^{n+1}, \dots, \widehat{w}_{\phi}^{N}\right]\right) \tag{4.10}$$

where $E^{n,j}\left(w_{\phi}^{n}\right)$ is a univariate objective function for the weight of the n-th feature at the j-th iteration.

As can be seen from Algorithm 1, first explicit and latent concepts of training queries are extracted from different sources (line 1) and then w_{ϕ} is randomly initialized (line 2). At each iteration of the proposed optimization method (line 3), w_{ϕ} is randomly shuffled (line 4). After that for each element of w_{ϕ} (line 5) and for each sampling policy (line 6), the objective function (i.e., $E^{n,j}\left(w_{\phi}^{n}\right)$ is sampled at the points $\boldsymbol{w}_{s,\phi}^{n} = [w_{\phi,m}^{n}]_{m=-M}^{M}$ (line 7). The sampling policy determines the values of M, K, and Δw at each iteration of the optimization approach. The smoothed objective function $\tilde{E}^{n,j}(w_{\phi,m}^{n})$ is obtained using the samples from $E^{n,j}\left(w_{\phi}^{n}\right)$ (line 7). Then, the optimum point of $\tilde{E}^{n,j}(w_{\phi,m}^{n})$ (i.e., $\widehat{w}_{\phi,m}^{n}$) is estimated (line 9). Next, the n-th element of w_{ϕ} is replaced by its estimated value (i.e., $\widehat{w}_{\phi,m}^{n}$) (line 10). These iterations continue until the number of iterations (i.e., j) goes beyond j_{max} (line 3) or convergence (lines 13-15).

Features

Table 5.4 summarizes all distinct features that are used to calculate the importance weight of each query concept c depending on its type. The list of concept types, which are determined by concept source, term representation and identification method, along with a set of features that are used to calculate the importance weight of query concepts of each type are

shown in Table 4.2. Concepts belonging to some concept types come from only one source, while other concept types assume two sources. For example, since the concepts of type TUU are UMLS concepts that are represented by unigrams and extracted from the top retrieved documents, this concept type is associated with two concept sources (top retrieved documents and UMLS).

As can be seen from Table 4.2, there are four different methods for identifying explicit and latent concepts in a query. The first and simplest method is to consider all unigrams and bigrams in a query or top retrieved documents as query concepts. The second approach uses MetaMap [6] to identify UMLS concepts in a query or top-retrieved documents. The third approach uses the Wikipedia-based health relatedness measure defined in [101] as:

$$hrm(c) = \frac{P(\ p \text{ is health} - \text{related} | c \in p)}{1 - P(p \text{ is health} - \text{related} | c \in p)}$$

Algorithm 1 Algorithm to optimize the feature weights with respect to the target retrieval metric using graduated optimization.

Identify explicit and latent concepts
Randomly initialize the feature weights vector (w_ϕ)
for $j = 1: j_{max}$ **do**
    Randomly shuffle w_ϕ
    for $n = 1 : N$ **do**
        for each sampling policy **do**
            Sample $E^{n,j}(w_\phi^n)$
            Obtain $\tilde{E}^{n,j}(w_{\phi,m}^n)$
            Obtain the optimum point $\widehat{w}_\phi^n$
            Update n-th element of $\boldsymbol{w}_\phi$ by $\widehat{w}_\phi^n$
        end for
    end for
    if Convergence **then**
        Break
    end if
end for

where $P(\ p \text{ is health} - \text{related}|c \in p)$ is the probability that a Wikipedia page p is health-related given that c occurs in p. Concepts for which this probability exceeds a pre-defined threshold are assumed to be health-related. The fourth approach uses the UMLS relationships table (MRREL.RRF table[4], which we further also refer to as the UMLS concept graph) to select the concepts related to the UMLS concepts identified in a query as latent concepts.

All features in Table 5.4, except Semantic Direction (SD), Semantic Popularity (SP) and Type Effectiveness (TE), are relatively simple and do not require a detailed explanation. Semantic direction is defined as follows. If S_c is the semantic type of concept c, So is the semantic type of the query concept o, to which concept c is related and $d(S_r, S)$ is the distance (i.e., the number of edges) from the root node (S_r) to node S in the UMLS semantic network, then the expansion concept c is defined to have an inward direction relative to the original concept o in the UMLS semantic network (i.e., the expansion concept is more general than the original query concept), if $d(S_r, S_c) < d(S_r;\ S_o)$. This feature is defined only for the UMLS expansion concepts that are related to the UMLS concepts in the original query.

Semantic popularity of concept c is defined as the number of concepts that are related to concept c in the UMLS concept graph (it can also be viewed as a node degree of concept c in the UMLS concept graph). A large value of this feature indicates popularity and generality of concept c. Type effectiveness is a binary feature that indicates whether the UMLS semantic type of concept c is effective for query expansion. As defined earlier, a semantic type is effective if its corresponding concepts can increase the precision of retrieval results when added to a query. The concept of effective semantic types for medical query expansion was first proposed in [55].

[4] http://www.ncbi.nlm.nih.gov/books/NBK9685/

Using the training queries and relevance judgments, we fine-tuned the set of effective semantic types from [55] to the collection and query sets used in this work. This will be explained in detail later.

4.3. Experiments

Experimental Setup

The experimental results reported in this work were obtained using the corpus, which includes around 730,000 documents from PubMed Central (PMC), and queries from the Clinical Decision Support (CDS) track at TREC 2014 [95] and 2015 [91]. 3-fold cross-validation was used to evaluate the performance of the proposed method (INTGR) and the baselines, which were first trained using the query set and relevance judgments from the CDS track of TREC 2014 to maximize infNDCG, the official retrieval metric of the CDS track [95]. The proposed method and the baselines were implemented using Indri retrieval toolkit[5]. The optimal values of Dirichlet prior, Jelinek-Mercer interpolation coefficient, the sizes of ordered and unordered bigram windows in the Indri query language were empirically determined to be 2500, 0:4, 4 and 17, respectively. Figure 4.3 illustrates how infNDCG changes by varying the number of PRF documents (used to extract concepts) and the number of concepts extracted from PRF documents. The values of these parameters that maximize infNCDG were used in experiments using TREC 2015 CDS track queries.

Besides the proposed graduated optimization approach, we used exhaustive line search to optimize individual feature weights as another baseline (INTGR-LS). This method examines the parameter space in uniform increments and chooses the setting that results in the highest

[5] http://www.lemurproject.org/indri/

Table 4.1: Brief description of features used to estimate the importance weight of concept c.

Feature	Description
TI	TF-IDF of concept c in the collection
CA	Average collection co-occurrence of concept c with other concepts in the query
CM	Maximum collection co-occurrence of concept c with other concepts in the query
NT	Number of top retrieved documents containing concept c
RS	Sum of retrieval scores of top-ranked documents containing concept c
TM	Maximum co-occurrence of concept c with other query concepts in top retrieved documents
TA	Average co-occurrence of concept c with other query concepts in top retrieved documents
GI	Do infoboxes of Wikipedia articles corresponding to concept c contain any health-related keywords?
IS	Does any of the terms of concept c exist in the title of any Wikipedia health-related articles?
CD	Average distance between concept c in the UMLS concept graph and other query, top document and related UMLS concepts identified for a query
SP	Popularity (node degree) of concept c in the UMLS concept graph
SD	Direction of concept c with respect to query concepts in the UMLS semantic network
TE	Does concept c have a UMLS semantic type that is effective for medical query expansion?

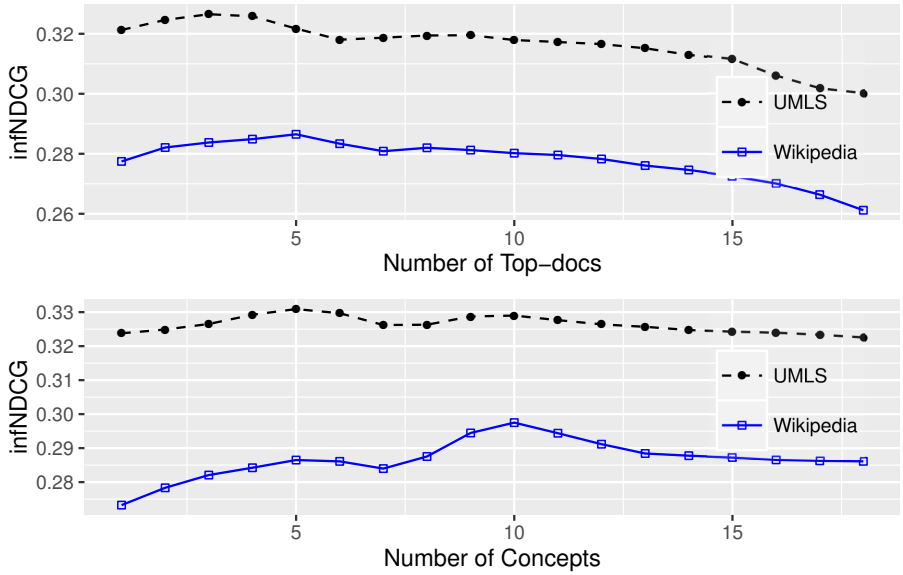


Figure 4.3: Average infNDCG on TREC 2014 CDS track queries by varying the number of top retrieved documents used to extract the concepts and the number of UMLS and Wikipedia concepts extracted from the top retrieved documents.

infNDCG. For both INTGR and INTGR-LS methods, the convergence threshold for the change in infNDCG was set to 0:001 and the number of iterations was limited to 20.

Baselines

The first baseline that was used in experiments is two-stage smoothing [124] (Two-Stage). Two-stage smoothing was also used as the smoothing method in implementing all other baselines and the proposed method. The other baselines used in experiments are Relevance Model (RM) [51], Parameterized Query Expansion (PQE) [17], Wiki-Orig and Wiki-TD [101], which use a Wikipedia-based health relatedness measure defined in (4.11). Other baselines that use only semantic concepts are UMLS-orig [94] and UMLS-TD [55]. UMLS-orig extracts UMLS concepts only from the query itself and breaks the phrases designating UMLS concepts into bigrams in order to incorporate them into the SDM retrieval model [70]. UMLS-TD extracts UMLS concepts from the top retrieved documents according to their semantic types. Since the original implementations of UMLS-TD and Wiki-TD are based on bag-of-words retrieval models, UMLS-TD_ and Wiki-TD_ are the modifications of UMLS-TD and Wiki-TD that use the SDM retrieval model to account for term dependencies when a concept is designated by a phrase.

We also compare the performance of the proposed method to the best performing methods (which used topic summaries as queries) in the CDS track of TREC in 2014 [77] and 2015 [13] (designated as TREC best). [77] used an ensemble of state-of-art unsupervised knowledge-based query expansion, re-ranking and relevance feedback methods. In [13], queries re expanded with unigrams and UMLS concepts identified in the query itself and the top retrieved documents.

Results

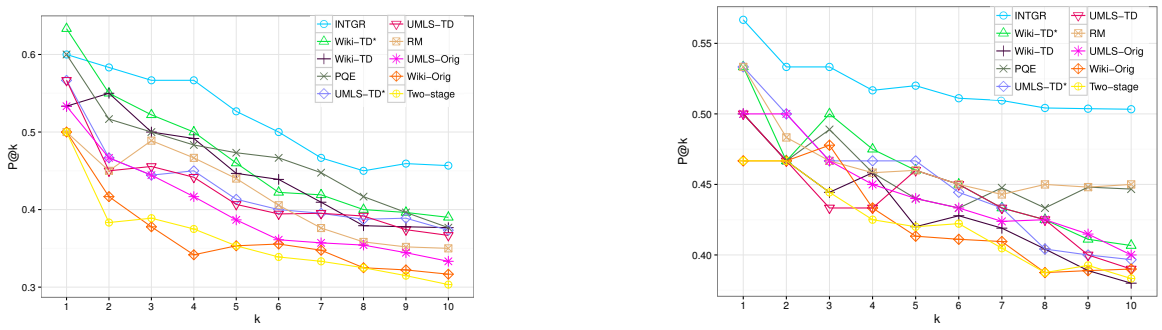
An initial list of 16 semantic types known to be effective for query expansion in medical records search was taken “as is” from [55]. We observed from the preliminary experiments that not all of these semantic types are effective for expansion of CDS queries. Therefore, we fine-tuned this initial list of semantic types by excluding those semantic types, for which the corresponding concepts did not improve infNDCG of retrieval results on training queries. The 5 semantic types retained from the initial list proposed in [55] are “Clinical Drug”, “Disease or Syndrome”, “Injury or Poisoning”, “Sign or Symptom” and “Therapeutic or Preventive Procedure”.

Tables 4.3 and 4.4 provide a summary of retrieval accuracy in terms of different retrieval metrics of the proposed method (INTGR) and the baselines on the query sets from the CDS track of TREC 2014 and 2015. As can be seen from Table 4.3, Wiki-TD* is the best performing baseline (since the best performing TREC methods are different for different query sets, they are not considered as the best performing baselines). Furthermore, the proposed algorithm outperforms INTGR-LS and the best methods in TREC 2014 and 2015.

Table 4.4 shows the degree of improvement and its statistical significance of the proposed method over the three best performing baselines (i.e., PQE, Wiki-TD, Wiki-TD*) and INTGR-LS. As follows from Table 4.4, INTGR significantly outperforms all of the best performing baselines in terms of all retrieval metrics. Using graduated non-convexity as a univariate optimization method results in 5-9% improvement of retrieval accuracy in terms of infNDCG, 10-23% improvement in terms of infAP and 8% improvement in terms of P@5 on different query sets.

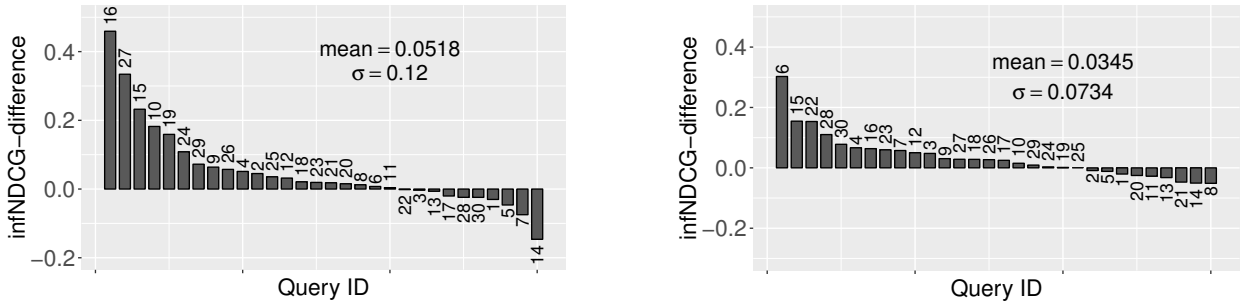
Table 4.5 illustrates the effect of using different knowledge bases in conjunction with INTGR on its performance in terms of different evaluation metrics. As follows from Table 4.5,

using INTGR only with Wikipedia results in the smallest improvement of retrieval accuracy across all retrieval metrics (and even a decrease of P@5). It also follows from this table that using INTGR with UMLS results in significantly greater improvement of all retrieval metrics, while the biggest improvement is achieved when explicit and latent concepts of a query are extracted from both UMLS and Wikipedia.



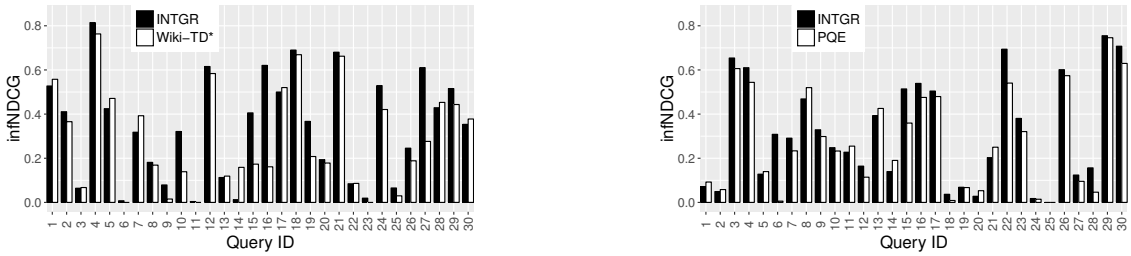
(a) TREC 2014 CDS track topics (b) TREC 2015 CDS track topics

Figure 4.4: Comparison of INTGR with the baselines in terms of P@k for $k \leq 10$ on the query sets from the CDS track of TREC 2014 and 2015.



(a) TREC 2014 CDS track topics (b) TREC 2015 CDS track topics

Figure 4.5: Topic-level differences of the infNDCG values for INTGR and the best-performing baselines (Wiki-TD* for TREC 2014 CDS track and PQE for TREC 2015 CDS track).



(a) TREC 2014 CDS track topics (b) TREC 2015 CDS track topics

Figure 4.6: Topic-level comparison of the infNDCG values for INTGR, the best performing baselines (Wiki-TD* for TREC 2014 CDS track and PQE for TREC 2015 CDS track).

Figure 4.4 provides performance comparison of INTGR with all of the baselines in terms of P@k for k from 1 to 10 (with a step size of 1). As can be seen from this figure, for all values of k except $k = 1$ in case of TREC 2014 CDS track queries, INTGR significantly outperforms all other baselines. It also follows from Figure 4.4 that for most of the values of k, the methods that expand the queries with the concepts extracted from the top-ranked documents (RM, UMLS-TD, UMLS-TD*, PQE, Wiki-TD, Wiki-TD* and INTGR) outperform the methods that represent the queries with the concepts extracted from them (Wiki-Orig and UMLS-orig). The average improvements of INTGR in terms of P@k for different values of k over the weakest and strongest baselines are 0:1560 and 0:0380, respectively, on the query set from TREC 2014 CDS track, while on the query set from TREC 2015 CDS track the improvements are 0:0988 and 0:0481, respectively.

Figure 4.5 illustrates topic level differences between the retrieval accuracy of INTGR in terms of infNDCG with the best performing baselines (Wiki-TD* for the CDS track of TREC 2014 and PQE for the CDS track of TREC 2015) on both query sets. From Figure 4.5(a), it follows that infNDCG of INTGR is greater than that of Wiki-TD* on 67% of the queries in the CDS track of TREC 2014, while from Figure 4.5(b) it follows that infNDCG of INTGR is greater than that of PQE on 73% of the queries in the CDS track of TREC 2015. The average improvement of INTGR over Wiki-

TD* in terms of infNDCG on TREC 2014 CDS track queries is 0.0518 with standard deviation 0.12, while the average improvement of INTGR over PQE in terms of infNDCG on TREC 2015 CDS track queries is 0.0345 with standard deviation 0.0734. The topics, on which INTGR has the greatest improvement and decline relative to Wiki-TD* in terms of infNDCG among those used in TREC 2014 CDS track are 16 (with 0.4593 improvement) and 14 (with 0.1462 decline). We can also observe that on the query set of TREC 2015 CDS track INTGR has the greatest improvement of 0.3026 and the greatest decline of 0.0512 in terms of infNDCG on topics 6 and 8, respectively. Figure 4.6 also provides a detailed comparison of retrieval accuracy of INTGR in terms of infNDCG with the best performing baselines (Wiki-TD_ for TREC 2014 CDS track and PQE for TREC 2015 CDS track) at the level of each individual topic in the CDS track of TREC 2014 and 2015.

We continued empirical evaluation of INTGR by analysis of its performance on difficult queries. We define a query as difficult if infNDCG of Two-Stage on this query is less than 0:1 and as very difficult if infNDCG of Two-Stage is less than 0.05. We observed that INTGR outperformed Wiki-TD* on 59% of difficult queries and on 86% of very difficult queries in the CDS track of TREC 2014. We also observed that INTGR outperformed PQE on 56% of difficult queries and on 77% of very difficult queries in CDS track of TREC 2014.

Discussion

Based on experimental analysis of INTGR presented in the previous section, we can conclude that the subset of UMLS semantic types that are effective for expansion of CDS queries is fairly small (includes less than 4% of UMLS semantic types). These semantic types can be grouped into three categories: “Disorders”, “Chemical & Drugs”, and “Procedures”. These three

categories in turn can be conceptually mapped to the three main types of CDS queries: “Diagnosis”, “Treatment”, and “Test”.

From tables 4.3 and 4.4, it follows that the proposed query representation method significantly outperforms all baselines in terms of all evaluation metrics and on both training and evaluation query sets. Furthermore, although INTGR was trained on the CDS track queries of TREC 2014 with the goal of maximizing infNDCG, INTGR also achieved significant (and, in many cases, even greater) improvement over the baselines in terms of other evaluation metrics (i.e., infAP and P@5) on both training and testing query sets. Also, as can be seen from Tables 4.3 and 4.4, the proposed method has significantly better performance when it is used in conjunction with graduated optimization method (INTGR) than when it is used with exhaustive line search (INTGR-LS), which we attribute to the ability of graduated optimization to efficiently find global optima of non-smooth and non-convex objective functions. Line search, on the other hand, may miss global optima, if the step size is not sufficiently small. In general, choosing the appropriate step-size is non-trivial and can dramatically affect the performance of line search.

As follows from Table 4.3, methods that utilize semantic (Wiki-TD/Wiki-TD* and UMLS-TD/UMLS-TD*) and statistical (RM and PQE) concepts for query representation and expansion behave differently on training and evaluation query sets. In particular, methods using semantic concepts show better results than the methods based on statistical concepts on the training query set, while the methods based on statistical concepts show better results on evaluation query set. However, the proposed method (INTGR) provides excellent results on both query sets, which indicates the utility of accounting for both types of concepts in a retrieval method for CDS queries. On the other hand, Table 4.5 demonstrates that for the methods based on semantic

Table 4.2. List of types for explicit and latent query concepts along with a set of features to estimate the importance of concepts of each type (Top-docs stands for top retrieved documents for the original query).

Concept Type	Concept Sources	Concept Representation	Concept Extraction	Features
QU	Query	unigrams	all query unigrams	TI, NT, RS, CA, CM, TA, TM
QOB	Query	ordered bigrams	all query bigrams	TI, NT, RS, CA, CM, TA, TM
QUB	Query	unordered bigrams	all query bigrams	TI, NT, RS, CA, CM, TA, TM
QUU	Query, UMLS	unigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM,TE, SP, CD
QUOB	Query, UMLS	ordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM,TE, SP, CD
QUUB	Query, UMLS	unordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
QDU	Query, Wikipedia	unigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
QDOB	Query, Wikipedia	Ordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
QDUB	Query, Wikipedia	unordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
TU	Top-docs	unigrams	direct identification	TI, NT, RS, CA, CM, TA, TM
TOB	Top-docs	ordered bigrams	direct identification	TI, NT, RS, CA, CM, TA, TM
TUB	Top-docs	unordered bigrams	direct identification	TI, NT, RS, CA, CM, TA, TM
TUU	Top-docs, UMLS	unigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
TUOB	Top-docs, UMLS	Ordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
TUUB	Top-docs, UMLS	unordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
TDU	Top-docs, Wikipedia	unigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
TDOB	Top-docs, Wikipedia	ordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
TDUB	Top-docs, Wikipedia	unordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
UU	UMLS	unigrams	UMLS relationships	TI, NT, RS, CA, CM, TA, TM, TE, SP, SD, CD
UOB	UMLS	ordered bigrams	UMLS relationships	TI, NT, RS, CA, CM, TA, TM, TE, SP, SD, CD
UUB	UMLS	unordered bigrams	UMLS relationships	TI, NT, RS, CA, CM, TA, TM, TE, SP, SD, CD

Table 4.3. Summary of retrieval accuracy of the proposed method and the baselines on the query sets from the CDS track of TREC 2014 and 2015.

Query set	**TREC 2014 CDS track**			**TREC 2015 CDS track**		
Method	**infNDCG**	**infAP**	**P@5**	**infNDCG**	**infAP**	**P@5**
Two-Stage [124]	0.1945	0.0493	0.3533	0.2110	0.0449	0.4200
Wiki-Orig [101]	0.2069	0.0550	0.3533	0.2193	0.0457	0.4133
UMLS-Orig [94]	0.2074	0.0569	0.3867	0.2206	0.0478	0.4400
RM [51]	0.2662	0.0836	0.4400	0.2765	0.0740	0.4600
UMLS-TD [55]	0.2577	0.1523	0.4067	0.2429	0.0748	0.4600
UMLS-TD*	0.2724	0.0810	0.4133	0.2503	0.0614	0.4667
PQE [17]	0.2796	0.0873	0.4733	0.2792	0.0762	0.4400
Wiki-TD [101]	0.2764	0.0881	0.4467	0.2418	0.0597	0.4267
Wiki-TD*	0.2883	0.0944	0.4600	0.2519	0.0633	0.4600
TREC best [77, 13]	0.2631	0.0757	0.4067	0.2928	0.0777	0.4467
INTGR-LS	0.3114	0.0993	0.4867	0.2987	0.0792	0.4800
INTGR	0.3401	0.1229	0.5267	0.3135	0.0873	0.5200

Table 4.4. Comparison of effectiveness of different knowledge bases on the query sets from the CDS track of TREC 2014 and 2015. Statistical significance and improvement in retrieval accuracy of the proposed method (INTGR) relative to its modification (INTGR-LS) and three best performing baselines (Wiki-TD, PQE and Wiki-TD*) on the query sets from the CDS track of TREC 2014 and 2015. * and † indicate statistically significant improvement with $p < 0.05$ and $p < 0.1$, respectively. Summary of retrieval accuracy of the proposed method and the baselines on the query sets from the CDS track of TREC 2014 and 2015.

Query set	**TREC 2014 CDS track**			**TREC 2015 CDS track**		
Method	**infNDCG**	**infAP**	**P@5**	**infNDCG**	**infAP**	**P@5**
Wiki-TD	23.05%*†	39.50%*†	17.91%*†	29.65%*†	46.23%*†	23.81%†
PQE	21.64%*†	40.78%*†	11.28%†	12.28%†	14.56%*	18.18%*†
Wiki-TD*	17.97%*†	30.19%*†	14.50%*†	24.45%*†	37.91%*†	13.04%*†
INTGR-LS	9.22%*†	23.77%*†	8.22%*†	4.95%*†	10.22%*	8.33%*†

concepts, UMLS is a better choice than Wikipedia with respect to all metrics, if only one knowledge repository is used. However, as follows from Table 4.5, combining both knowledge bases results in better retrieval accuracy than using any one of them individually. Although from Figures 4.4 and 4.5 as well as Tables 4.3 and 4.4 it follows that INTGR has slightly lower accuracy improvement over its best-performing baseline and Two-Stage on the testing query set than on the training query set, the improvement that INTGR achieves over Two-Stage is much higher than the improvement of the best performing baseline over Two-Stage. However, as follows from Figures 4.6 and 4.5, there is a greater number of topics on which INTGR has better retrieval accuracy than the best performing baseline on both training and testing query sets. Therefore, based on these observations, we can conclude that INTGR is robust to overfitting, due to its use of multiple and diverse relevance signals and concept sources.

Table 4.5. Comparison of effectiveness of different knowledge bases on the query sets from the CDS track of TREC 2014 and 2015.

Query set	**TREC 2014 CDS track**			**TREC 2015 CDS track**		
Method	**infNDCG**	**infAP**	**P@5**	**infNDCG**	**infAP**	**P@5**
INTGR using no knowledge bases	0.2673	0.0875	0.4601	0.2771	0.0758	0.4633
INTGR using only Wikipedia	0.2975 (11.30%)	0.0936 (6.97%)	0.4533 (-1.47%)	0.2954 (6.60%)	0.0779 (2.77%)	0.4667 (0.09%)
INTGR using only UMLS	0.3309 (23.79%)	0.1170 (33.71%)	0.5200 (13.02%)	0.3012 (8.67%)	0.0786 (3.93%)	0.5033 (7.93%)
INTGR using UMLS and Wikipedia	0.3401 27.23%)	0.1229 (40.46%)	0.5267 (14.47%)	0.3135 (13.14%)	0.0873 (15.17%)	0.5200 (11.52%)

CHAPTER 5 A SEQUENTIAL APPROACH TO EXTRACT EXPANSION CONCEPTS FROM CONCEPT GRAPHS

4.1. Introduction

Concept graphs can be constructed manually (e.g. ConceptNet [57]), or automatically from a given collection [3, 11, 47, 48] by considering any pair of terms or phrases that frequently co-occur in the same context (e.g., document) as semantically related. Concept graphs are utilized for query expansion by selecting the concepts related to the ones occurring in the query. However, since concept graphs are typically dense [57], there can be a large number of concepts that are immediately related to the query concepts. Although it has been previously shown that there exist very effective expansion concepts in remote layers of concepts related to the original query concepts (i.e., concepts with one or more intermediate concepts between them and the query concepts) [48], the number of candidate concepts that need to be evaluated increases exponentially with the number of layers to consider. However, only a small fraction of hundreds or potentially thousands of concepts that can be discovered in all layers of related concepts in the concept graph can improve retrieval results, while others need to be discarded to avoid noise and concept drift [50, 74, 84]. Figure 5.1 illustrates this problem for the query "*poach preserve wildlife*", which we will use as an example throughout this work. According to ConceptNet 5, there are 374 concepts in the first layer of related concepts (that are directly related to the query concepts). Some of these concepts, such as "hunt" and "nature preserve", are relevant to the information need behind this query and are useful expansion concepts. However, other related concepts, such as "boil", "injure", "keep", "album" are not relevant to the information need behind this query and should be discarded. The concepts in the third layer, such as "capture" and

"wildlife sanctuary" that are also related to the information need behind this query should be separated from many other non-relevant concepts in this layer, even though some of these non-relevant concepts are related to the useful concepts in the second layer.

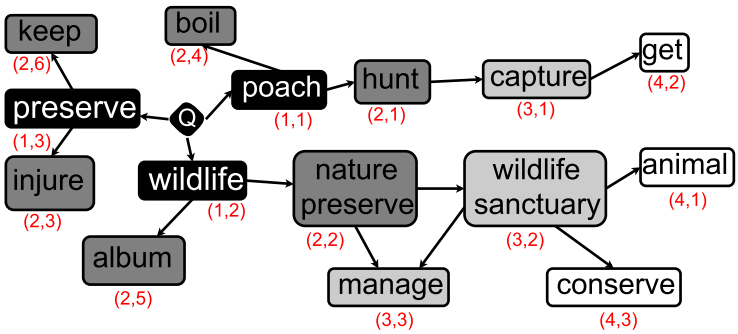
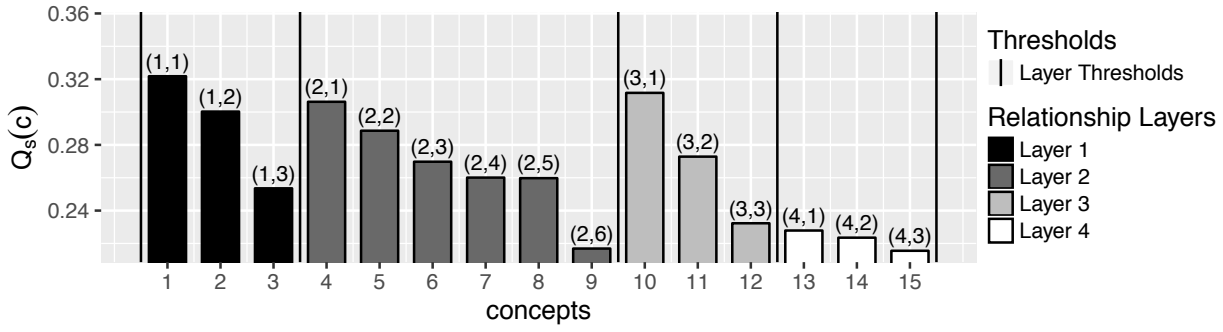


Figure 5.1: Fragment of the concept graph of ConceptNet 5 showing the concepts related to the concepts in the query "poach wildlife preserve". The first number in parenthesis indicates concept layer, the second number is the index of a concept in the concept layer.

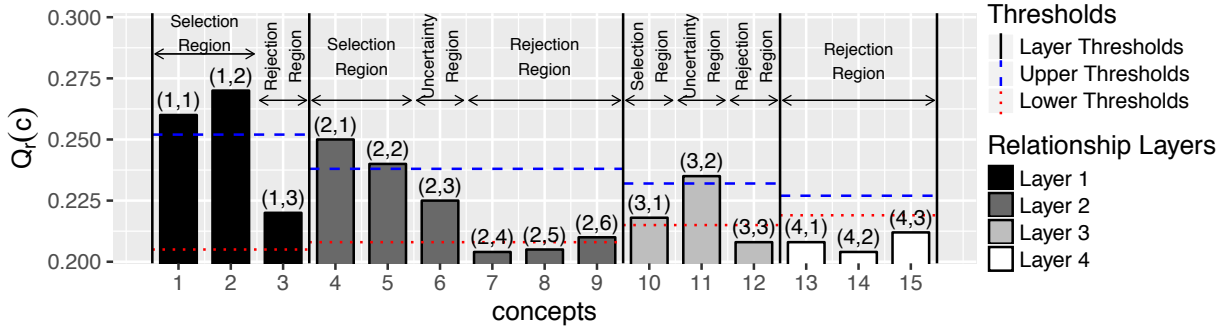
Therefore, accurate evaluation and effective pruning of noisy concepts to find a small number of highly effective concepts for query expansion are the two fundamental challenges in effective utilization of concept graphs for query expansion. In this paper, we propose a two-stage method that addresses these challenges. The proposed method is illustrated for the case of our example query in Figure 5.2.

In the first stage of the proposed method, all concepts in each concept layer are first sorted according to a quality measure calculated using a number of computationally inexpensive features, such as TF-IDF. Then, in the second stage of the method, a concept selection method that relies on more computationally expensive features is applied to *sequentially* select a set of expansion concepts from the concepts in each layer that are sorted in the first stage. This method selects the concepts from each layer in a one-by-one manner while *maintaining the desired level of precision and minimizing the number of concepts that need to be examined*. Therefore, a

limited number of concepts are examined in each layer using computationally expensive features and a limited number of them are selected as expansion concepts. To improve the efficiency and avoid topic drift, only the concepts that are related to the concept selected in layer i are considered in layer $i + 1$. As a result, the proposed method avoids calculating computationally expensive features, such as average mutual information, for a large number of concepts in concept layers that are further away from the original query concepts.



(a) Initial Concept Sorting (Stage I).



(b) Sequential Concept Selection (Stage II)

Figure 5.2: Illustration of the proposed two-step concept selection method for a set of related concepts in Figure 5.1.

5.2. Method

Due to a large number of candidate concepts that are related to the original query concepts, finding effective expansion concepts in a concept graph is a challenging problem,

particularly since most of the candidate concepts have zero or negative effect on the accuracy of retrieval results, when they are used for query expansion. The proposed query expansion method is based on the idea of sequential examination of concepts in different layers of a concept graph with respect to the original query concepts. It first evaluates the related concepts at each relationship layer by using a number of inexpensive features and then chooses subsets of related concepts to be evaluated carefully by using more expensive features. The method aims to minimize the total number of concepts evaluated in each layer, while maintaining the precision of retrieval results above a given threshold. This way, selection of effective expansion concepts can be formulated as an optimization problem, in which the objective is to minimize the total number of evaluated concepts subject to precision of retrieval results being above a given threshold.

In this section, we present the details of our proposed method to address the problem of selection of effective expansion concepts from dense, large and noisy concept graphs. First, we discuss the details of the adopted query expansion model and then present the methods to construct concept graphs and use them for sequential selection of query expansion concepts.

Query Expansion

The proposed method is based on the Latent Concept Expansion (LCE) [70] framework. LCE was designed to incorporate the query expansion terms from the top retrieved documents into Markov Random Fields-based retrieval models [69], which allow to account for term dependencies. The proposed method uses the following scoring function of document D with respect to query Q:

$$s(Q, D) = \sum_{i=0}^{k} \alpha_i \sum_{j=1}^{M_i} f_i(D, C_{(i,j)}) \tag{5.1}$$

where α_i is the weight of the concepts in the i-th concept layer, k is the number of concept layers that are involved in the concept selection process, and M_i is the number of concepts in the i-th concept layer. $C_{(i,j)}$ in the above equation is the j-th concept in the i-th concept layer. Let us define $\mathbb{C}_i = \{C_{(i,j)}\}_{j=0}^{M_i}$ as the set of concepts in the i-th concept layer. In this case, $\mathbb{C}_0$contains all the unigrams in a given query. Retrieval models using unigrams only utilize $\mathbb{C}_0$. $\mathbb{C}_1$ includes the query concepts that can be found in the concept graph.

A query is expanded with a limited number of concepts selected in each concept layer $1 \leq i \leq k$. In the above formula, $f_i(D, C_{(i,j)})$ is the matching score of concept $C_{(i,j)}$ in document D. Let us define

$$g(k, D) = \log\left(\frac{tf_{k,D} + \mu\frac{cf_k}{|C|}}{|D| + \mu}\right) \tag{5.2}$$

as the matching score of concept k with respect to document D. In the above equation, $g(k, D)$ is the log-likelihood of k in the language model of D smoothed using Dirichlet prior smoothing, μ is the Dirichlet prior, $|D|$ is the length of document D and $|C|$ is the number of documents in a collection. k can be a unigram w, ordered $\#uw(b)$ or unordered $\#od(b)$ bigram b. Any other n-gram concepts are represented in terms of these three concept types. For example, the concept "*wild life preserve*" is decomposed into a set of unigrams ("*wild*", "*life*", "*preserve*") and a set of bigrams ("*wild life*", "*life preserve*"). Therefore, the matching score of document *D* with respect to concept $C_{(i,j)}$is defined as:

$$f_i(D, C_{(i,j)}) = \gamma T \sum_{w \in C_{(i,j)}} g(w, D) + \gamma U \sum_{b \in C_{(i,j)}} g(\#uw(b), D)$$

$$+ \gamma O \sum_{b \in C_{(i,j)}} g(\#od(b), D) \quad (5.3)$$

where γT, γO, and γU are the weights of unigrams, ordered and unordered bigrams, respectively. By replacing Dirichlet smoothing in (5.2) with Jelinek-Mercer smoothing and considering only the concepts from the top retrieved documents as expansion concepts, we obtain the same retrieval function as used in the original LCE model [70].

The proposed method for query expansion consists of two stages. In the first stage, candidate expansion concepts are ordered with respect to a quality measure (defined below), while a sequential selection method to find the expansion concepts is applied in the second stage. As a result, only the concepts that are likely to be useful expansion concepts are evaluated in detail. Therefore, the key idea behind the proposed method is to use computationally inexpensive features to initially sort all related concepts and a combination of computationally expensive and inexpensive features to sequentially evaluate them and select the final set of concepts for query expansion. Sorting of the concepts in Stage I of the proposed method provides an initial understanding of concept usefulness, which is utilized in Stage II to minimize the number of evaluated concepts. These two stages as well as different methods to construct the concept graph are explained in more detail below.

Concept Graphs

Concept graphs used in experiments were constructed in two different ways. One way is to use a manually created semantic network, such as ConceptNet [57]. In this case, we only considered English concepts. If there is a link between the two concepts in ConceptNet, they are considered as related concepts in the concept graph. The other way to construct a concept graph is to use a collection itself [47]. Only unigram concepts are used in the concept graph in this case.

We used Hyper-space Analogue to Language (HAL) similarity measure [24] as a measure of semantic relatedness between the concepts. HAL considers two concepts as highly related if they frequently appear together within a sliding window of certain size (typically, 20 words) throughout a given document collection.

Sequential Concept Expansion

When concept graphs are large and dense, a very large number of concepts needs to be evaluated to select the useful expansion concepts. If we define $\mathbb{C}^u$ as the set of useful concepts (i.e., those that increase the precision of retrieval results, if added to a query) and $\mathbb{C}$ as the set of all concepts in a concept graph, then the optimal solution to the concept selection problem is obtained by examining all possible subsets of expansion concepts with size 0 to $|\mathbb{C}|$. To obtain this optimal solution, $2^{|\mathbb{C}|}$ subsets of concepts should be evaluated, which is clearly infeasible for any meaningful number of concepts.

A simplified suboptimal solution for the concept selection problem is to evaluate only the concepts that are directly related to the query concepts via a number of intermediate concepts. To further simplify the concept selection process, instead of exhaustively examining all related concepts, we propose to evaluate them sequentially (i.e., one after the other). In this approach, starting from the query concepts, the concepts in closer concept layers (i.e., the ones that are semantically closer to the query concepts) are evaluated first. Although the concepts that are semantically closer to the query concepts are not necessarily more useful concepts, they are less affected by the noise propagated from the other concept layers.

Let us define $\mathbb{C}^r_{(i,j)}$ and $\mathbb{C}^u_{(i,j)}$ as the sets of *related* and *useful* concepts, respectively, when examining $C(i,j)$, the j-th concept at relationship level i. Selection of the concept $C(i,j)$ for

query expansion can be formulated as a binary hypothesis testing problem with the null hypothesis H_0 and an alternative hypothesis H_1 defined as follows:

$$\mathrm{H}_0\colon\ C_{(i,j)} \in \mathbb{C}^r_{(i,j)} - \mathbb{C}^u_{(i,j)}$$

$$\text{v. s.} \qquad \mathrm{H}_1\colon C_{(i,j)} \in \mathbb{C}^u_{(i,j)} \tag{5.4}$$

After a concept is selected from $\mathbb{C}^r_{(i,j)}$, it is removed from this set. Selecting a concept and adding it to the query changes the usefulness of other concepts; therefore $\mathbb{C}^u_{(i,j)}$ should also be modified after a concept is selected for query expansion.

Stage I: Initial Sorting of Concepts

The concepts are first sorted according to a linear combination of computationally inexpensive features:

$$\tilde{Q}_s(c) = \sum_{j=1}^{m_s} \bar{\lambda}_{s,j} f_j(c), \tag{5.5}$$

where $\tilde{Q}_s(c)$ is a quality measure of concept c, $f_j(c)$ is a feature function, $\bar{\lambda}_{s,j}$ is a feature weight, and m_s is the number of inexpensive features.

Stage II: Sequential Selection of Concepts

Let us define $\tilde{\mathbb{C}}^u_i$ as the set of concepts selected in the concept layer$i \in \{1, 2, \dots, k\}$. It is preferable for the set $\tilde{\mathbb{C}}^u_i$ to be as close as possible to the set of useful concepts in the concept layer i (i.e., $\mathbb{C}^u_i$). In each concept layer starting from the first (i.e., $C_{(i,1)}$), the concepts are evaluated sequentially. After examining the k-th concept layer, the total set of selected concepts is the union the concepts selected in each of the $\{1, 2, \dots, k\}$ concept layers:

$$\tilde{\mathbb{C}}^{ut}_k = \bigcup_{i=1}^{k} \tilde{\mathbb{C}}^u_k \tag{5.6}$$

An entire set of selected concepts can be obtained by solving the following optimization problem:

$$\min_{\tilde{\mathbb{C}}_k^{ut}} \left\{ \sum_{i=1}^{k} N_i \right\}$$

$$\text{such that } E\left(\widetilde{\mathbb{R}}_\Lambda; \mathbb{T}\right) > \theta_Q \tag{5.7}$$

In the above equation, N_i is the number of concepts evaluated in the i-th concept layer. N_i is less than or equal to the number of concepts in the i-th concept layer (i.e., $N_i \leq M_i$). $E\left(\widetilde{\mathbb{R}}_\Lambda; \mathbb{T}\right)$ is a retrieval quality evaluation metric for a set of document rankings, $\widetilde{\mathbb{R}}_\Lambda$, based on the training data $\mathbb{T}$. Document rankings $\widetilde{\mathbb{R}}_\Lambda$ are those that correspond to the expanded query, which contains the selected concepts $\tilde{\mathbb{C}}_k^{ut}$. In (5.7), θ_Q is a pre-specified lower threshold for $E\left(\widetilde{\mathbb{R}}_\Lambda; \mathbb{T}\right)$.

The goal of the above optimization procedure is to address the problem of dealing with a large number of related concepts that need to be evaluated in each concept layer. This goal is accomplished by *selecting* concepts in such a way that the least number of concepts is evaluated, while maintaining an acceptable value for the target retrieval metric (e.g. MAP). The set $\tilde{\mathbb{C}}_k^{ut}$ can be approximated by Algorithm 2. In this algorithm, $\tilde{Q}_r\left(C_{(i,j)}\right)$ is a measure of retrieval effectiveness of the candidate concept $C_{(i,j)}$ that can be calculated using expensive and inexpensive features as a weighted linear combination of feature functions as follows:

$$\tilde{Q}_r\left(C_{(i,j)}\right) = \sum_{j=1}^{m_r} \bar{\lambda}_{r,j} f_j\left(C_{(i,j)}\right), \tag{5.8}$$

where $\bar{\lambda}_{r,j}$ is the weight of a feature function $f_j\left(C_{(i,j)}\right)$, and m_r is the number of expensive and inexpensive features. $\tilde{Q}_r\left(C_{(i,j)}\right)$ is applied to the concepts that are already sorted using $\tilde{Q}_s(c)$. Different decisions can be made by comparing $\tilde{Q}_r\left(C_{(i,j)}\right)$ with the upper and lower thresholds (denoted by β_U and β_L). One of the decisions that can be made as a result of such comparisons is whether to select $C_{(i,j)}$ as an expansion concept or to discard it. The other decision is whether

to continue examining and evaluating the concepts in the same concept layer or to switch to the next concept layer and start examining its concepts. These decisions are formalized in Table 5.1.

Computational complexity of this algorithm can be reduced further by discarding the concepts that have $Q_s(c)$ below a threshold $\beta_{s,L}$ in stage I of the algorithm (i.e., those that have $Q_s(c) < \beta_{s,L}$). In this case, the number of concepts that are evaluated in the Stage II of the algorithm can be decreased at the expense of retrieval performance degradation, the degree of which is controlled by the value of β_L.

5.3. Experiments

Statistics of the collections used for experimental evaluation of the proposed method are shown in Table 5.2. Parameters and hyperparameters of the proposed method and the baselines were optimized with respect to the Mean Average Precision (MAP) on the training set. The concepts in the first concept layer are obtained by using different methods depending on how the concept graph was constructed. If the concept graph is constructed from the collection, this set of concepts consists of all unigrams in the query. If the concept graph is obtained from ConceptNet, this set of concepts consist of the longest query n-grams that correspond to

Table 5.1. Three possible decisions that can made by evaluating concept c using the proposed method.

Decision	**Criterion**
Select concept $C_{(i,j)}$ & continue with the same concept layer	If $\tilde{Q}_r\left(C_{(i,j)}\right) \geq \beta_U$
Discard concept $C_{(i,j)}$ & continue with the same concept layer	If $\beta_L \leq \tilde{Q}_r\left(C_{(i,j)}\right) < \beta_U$
Discard concept $C_{(i,j)}$ & move to the next concept layer	If $\tilde{Q}_r\left(C_{(i,j)}\right) < \beta_L$

Algorithm 2 The proposed two-stage algorithm to obtain a set of expansion concepts.

$i = 1$
$\tilde{\mathbb{C}}_k^{ut} = \{\}$
do
    $\tilde{\mathbb{C}}_i^{u} = \{\}$
    for $c \in \mathbb{C}_i$ **do**
        compute $\tilde{Q}_s(c)$
    end for
    sort $\mathbb{C}_i$ according to $\tilde{Q}_s(c)$
    for $j = \{1, \dots, M_i\}$ **do**
        compute $\tilde{Q}_r\left(C_{(i,j)}\right)$
        if $\tilde{Q}_r\left(C_{(i,j)}\right) < \beta_U$ **then**
            add $C_{(i,j)}$ to $\tilde{\mathbb{C}}_i^{u}$
        end if
        if $\tilde{Q}_r\left(C_{(i,j)}\right) < \beta_L$ **then**
            $i = i + 1$
        end if
    end for
    $\tilde{\mathbb{C}}_k^{ut} = \tilde{\mathbb{C}}_k^{ut} \cup \tilde{\mathbb{C}}_i^{u}$
while $\tilde{\mathbb{C}}_i^{u} \neq \{\}$

Table 5.2. Statistics of experimental collections.

Collection	# of documents	# of terms
TREC7-8	472,526	2.16×10^8
ROBUST04	528,155	2.53×10^8
GOV	1,247,753	1.37×10^9

ConceptNet concepts. The concepts in other concept layers were selected by using the links between the concepts in the constructed concept graph, and they can be n-gram concepts with $n \geq 1$.

Baselines

The primary goal of the sequential concept selection method presented in Section 5.2 is to minimize the number of evaluated candidate expansion concepts from the concept graph. Considering the trade-off between the precision and the computation time, four variations of the

proposed method, which are summarized in Table 5.3 and Figure 5.3, are considered as baselines in experiments. In Table 5.3:

$$Q_b\left(C_{(i,j)}\right) = \sum_{j=1}^{m_s} \hat{\lambda}_{b,j} f_j\left(C_{(i,j)}\right) \tag{5.9}$$

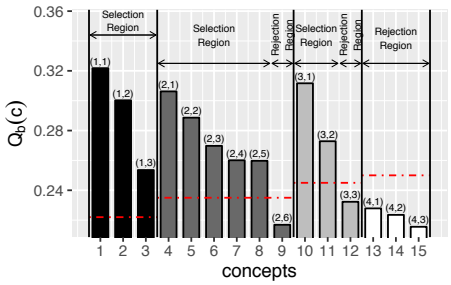
is a quality measure computed as a linear weighted combination of the feature functions. It is assumed that the retrieval system has limitation on computational complexity. So, the set of features used to calculate the quality measure $Q_b(c)$ for the baselines is the same as the set of features used to calculate $Q_s(c)$ in (5.5) for our proposed method. In Table 5.3, $I(c)$ is the index of a concept in the sorted set of concepts and Li is the number of selected concepts from the i-th concept layer.

Table 5.3. Summary of the proposed method and the baselines

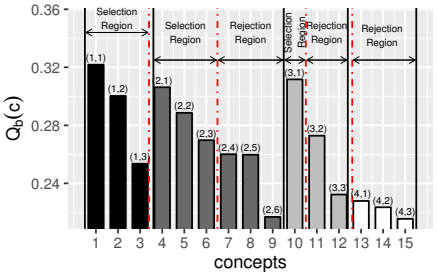
Method	Optimization Problem		Criteria in the Approximate Solution		
	Objective	**Constraint**	**Selecting**	**Rejecting**	**Stopping**
Method A	$\min\{\sum_{i=0}^{k} L_i\}$	$E(\widetilde{\mathbb{R}}_{\Lambda}^{k}; \mathbb{T}) > \theta$	$Q_b(c) > \beta_Q$	$Q_b(c) < \beta_Q$	$i > k$
Method B	$\max\{E(\widetilde{\mathbb{R}}_{\Lambda}^{k}; \mathbb{T})\}$	$\sum_{i=0}^{k} L_i < \theta$	$I_i(c) < \beta_I$	$I_i(c) > \beta_I$	$i > k$
Method C	$\min\left\{\sum_{i=0}^{k} L_i\right\}$	$E(\widetilde{\mathbb{R}}_{\Lambda}^{k}; \mathbb{T}) > \theta$	$Q_b(c) > \beta_Q$	$Q_b(c) < \beta_Q$	$i > k$
Method D [48]	$\max\{E(\widetilde{\mathbb{R}}_{\Lambda}^{k}; \mathbb{T})\}$	$\sum_{i=0}^{k} L_i < \theta$	$I(c) < \beta_I$	$I(c) > \beta_I$	$i > k$
Proposed	$\min\left\{\sum_{i=0}^{k} N_i\right\}$	$E(\widetilde{\mathbb{R}}_{\Lambda}^{k}; \mathbb{T}) > \theta$	$Q_r(c) > \beta_U$	$Q_r(c) < \beta_L$	$L_i = 0$

As follows from Table 5.3, when expansion concept selection problem is formulated as minimization of the number of concepts by keeping the evaluation metric above a desired level (i.e., methods A and C), the approximate solution is to select concepts if their quality measure $Q_b(c)$ is above a threshold and reject otherwise. But, in the case of maximization of retrieval precision by putting a constraint on the number of selected concepts (i.e., methods B and D), the

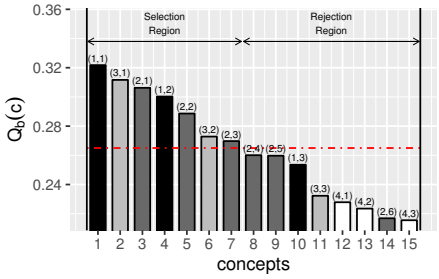
approximate solution is to select a limited number of concepts that result in the highest improvement in average precision.



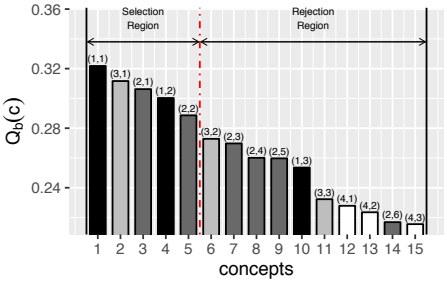
(a) Method A: Single threshold on $Q_b(c)$ in each layer



(b) Method B: Single threshold on $I(c)$ in each layer.



(c) Method C: Single threshold on $Q_b(c)$ in all layers



(d) Method D: Single threshold on $I(c)$ in all layers [48]

Figure 5.3: Graphical summary of the baselines A-D. The thresholds placed on the quality of concepts ($Q_b(c)$) or the number of selected concepts ($I(c)$) in each or all of the concept layers are shown by the red lines.

As can be seen from Table 5.3, methods A and B, similar to our proposed method, but unlike methods C and D, select the expansion concepts from different relationship layers sequentially. In other words, in methods A and B and in our proposed method, the concepts in concept layer i are examined, if their ancestor concept nodes in concept layer $i-1$ are selected. However, methods C and D first find a set of all concepts in the layers $1 \leq i \leq k$ and examine all of them at once. Since these methods do not prune the concepts in previous concept layers, noise can get propagated from layer $i-1$ to layer i.

In methods B and D, the threshold (indicated by β_i in Table 5.3) is on the number of selected concepts, but, in methods A and C and the proposed method, the thresholds (shown by β_Q, β_L , and β_U in Table 5.3) is on the quality of concepts. Therefore, unlike methods B and D, the thresholds in methods A and C and the proposed method do not limit the number of expansion concepts, and depending on the query, the collection and the required level of retrieval accuracy, the optimal number of expansion concepts is determined automatically. Although methods A and C and the proposed method do not use a predefined threshold on the number of expansion concepts, they use a predefined threshold on quality measures ($Q_b(c)$ or $Q_r(c)$). In methods A and B and the proposed method, there are distinct thresholds for each concept layer, while in methods C and D, there is only one single threshold for all concept layers. As described in more detail later, β_Q and β_I as well as β_L and β_U are optimized with respect to their objective functions and constraints by using a coordinate descent method.

Our proposed method stops at the concept layer i, if no concept is identified at this layer (i.e., if $L_i = 0$), but the methods A-D have predefined limits on the total number of examined concept layers (i.e., k). In other words, the proposed method stops when there is not enough evidence that there are useful concepts in other concept layers, while methods A-D stop when they examine a given number of concept layers. Therefore, unlike the baselines A-D, the number of concept layers examined by the proposed method differ from query to query.

Finally, none of the baselines A-D consider minimizing the number of evaluated concepts. The constraints used by methods B and D are on the number of selected concepts, and the objective functions of methods A and C are minimizing the total number of selected concepts.

Table 5.4. Features used in stages I and II of the proposed method. All of the listed features are considered in stage II of the proposed method, but only the features without asterisks are considered in Step I of the proposed method.

Feature	Description
hgstDocScore	Retrieval score of the highest ranked document containing $C_{(i,j)}$
avgDocScore	Average retrieval score of all documents containing $C_{(i,j)}$
varDocScore	Variance of retrieval score of all documents containing $C_{(i,j)}$
avgTDocScore	Average retrieval scores of the top documents containing $C_{(i,j)}$
termFreqTpDoc	Number of occurrences of $C_{(i,j)}$ in the top documents
docFreqTpDoc	Number of top documents containing $C_{(i,j)}$
nodeDegree	Node degree of $C_{(i,j)}$ in the concept graph
avgNumLinks	Average number of paths between $C_{(i,j)}$ and query concepts
maxNumLinks	Maximum number of paths between $C_{(i,j)}$and query concepts
avgCooccur*	Average co-occurrence of $C_{(i,j)}$ with query concepts
maxCooccur*	Maximum co-occurrence of $C_{(i,j)}$ with query concepts
maxTCooccur	Maximum co-occurrence of $C_{(i,j)}$ with query concepts in top retrieved documents
avgTCooccur	Average co-occurrence of $C_{(i,j)}$ with query concepts in top retrieved documents
avgTCooccurP*	Average co-occurrence of $C_{(i,j)}$ with at least a pair of query concepts in top retrieved documents
maxTCooccurP*	Maximum co-occurrence of $C_{(i,j)}$ with at least a pair of query concepts in top retrieved documents
avgTCooccur*	Average co-occurrence of $C_{(i,j)}$ with all previously selected concepts in top retrieved documents
maxTCooccur*	Maximum co-occurrence of $C_{(i,j)}$ with all previously selected concepts in top retrieved documents
avgCooccurL*	Average co-occurrence of $C_{(i,j)}$ with selected concepts in concept layer $i-1$
maxCooccurL*	Maximum co-occurrence of $C_{(i,j)}$ with selected concepts in concept layer $i-1$
avgTCooccurL*	Average co-occurrence of $C_{(i,j)}$ with selected concepts in concept layer $i-1$ in top retrieved documents
maxTCooccurL*	Maximum co-occurrence of $C_{(i,j)}$ with selected concepts in concept layer $i-1$ in top retrieved documents
avgTMiP*	Average mutual information of $C_{(i,j)}$ with at least a pair of query concepts in top retrieved documents
maxTMiP*	Maximum mutual information of $C_{(i,j)}$ with at least a pair of query concepts in top retrieved documents
avgTMiL*	Average mutual information of $C_{(i,j)}$ with selected concepts in concept layer $i-1$ in top retrieved documents
maxTMiL*	Maximum mutual information of $C_{(i,j)}$ with selected concepts in concept layer $i-1$ in top retrieved documents

The other baselines that are considered in our experimental evaluation are Query Likelihood retrieval model [86] with Dirichlet prior smoothing (QL) [122], Relevance Model (RM) [51], Sequential Dependence Model (SDM) [69] and Latent Concept Expansion (LCE) [70].

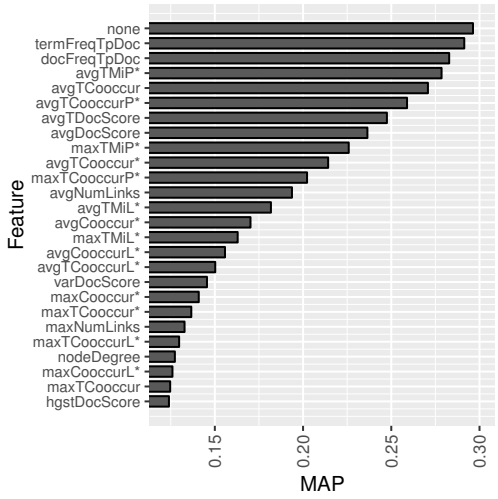


Figure 5.4: MAP after removing one feature from the list of features in Table 5.4 that results in the highest decrease of MAP at a time.

Features

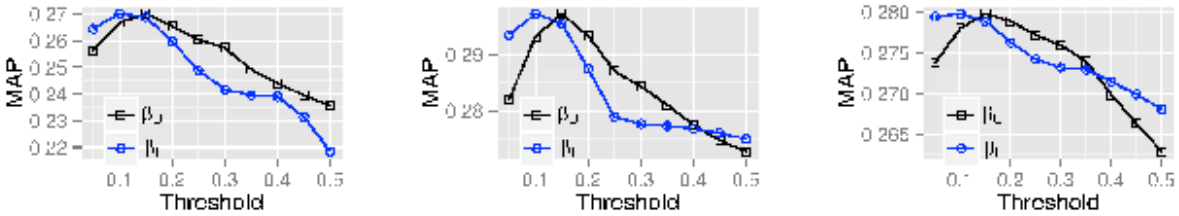
Two sets of features are used in the proposed two-stage method. The first set consists of only computationally inexpensive features that are used to initially sort the concepts in the first stage of the proposed method. The second set consists of mostly computationally expensive features that are used to select the concepts in the second stage of the proposed method. Computationally expensive features include the ones that are based on co-occurrence and mutual information [64]. Specifically, the first set of features is used to calculate $Q_s(C_{(i;j)})$ in (5.5) and the second set is used to calculate $Q_r(C_{(i;j)})$ in (5.8).

According to Table 5.4, the number of inexpensive features (designated by m_s in (5.5)) is 11, and the total number of expensive and inexpensive features (designated by m_r in (5.8)) is 25. In this table, 16 features depend on the top retrieved documents, 6 on the collection and 3 on

the concept graph. The top retrieved documents are obtained only once using SDM retrieval model with the original query. The number of top retrieved documents is a hyper-parameter of the proposed method that is estimated using cross-validation.

To determine the relative importance of features, we conducted a study, the results of which for the ROBUST04 collection are reported in Figure 5.4. In this study, we started with a full feature set and removed one feature, which results in the highest reduction of MAP after being removed from the feature set, at a time. The weights of other features have been updated to satisfy the conditions of the optimization problem each time a feature was removed. As follows from Figure 5.4, the features that are utilized in both stages of the proposed method have the highest impact on its retrieval accuracy. It can be also concluded that the features that are dependent on the collection tend to have a greater effect on retrieval performance than other features. Finally, when all the features are removed, retrieval results are obtained using only the concepts in the original query, which have a higher importance weight relative to the expansion concepts.

Different combinations of the features listed in Table 5.4 can be utilized for query expansion, depending on the collection and query set. In particular, from an entire set of features listed in Table 5.4 we obtained smaller sets of highly effective features for each experimental collection via a backward feature elimination process, when the features that have negative effect on retrieval accuracy are eliminated one at a time.



(a) TREC 7-8 (b) ROBUST04 (c) GOV

Figure 5.5: MAP of the proposed method in terms of β_U and β_L at the 2nd concept layer.

Three-fold cross validation was used to evaluate the performance of the proposed method and the baselines. At each cross validation fold, the thresholds β_U and β_L for each concept layer as well as the weights of the features in stages I and II of the proposed method (i.e., $\lambda_{s.j}$ and $\lambda_{r.j}$ in (5.5) and (5.8)) were optimized in such a way that the MAP of the top retrieved documents stays above the threshold θ, while the number of concepts examined in stage II of the proposed method is minimized. Coordinate ascent [71] was used to optimize the values of these parameters. Starting from an initial random point, the parameter space was examined in uniform steps (step size was 0.01), one parameter at a time. This process was repeated for all parameters until convergence (if the change in the target retrieval metric from one iteration to another is less than 0.05) or until the number of iterations exceeds 100. The values of θ were chosen based on the MAP of retrieval results of the QL method. The values of θ for TREC 7-8, ROBUST04, and GOV collections were set to 0.28, 0.32, and 0.30, respectively, all of which are greater than the MAP of the QL method by 0.08 (see Table 5.6).

The same training procedure with the same θ as above was used to optimize the parameters of the baseline methods, such as $\hat{\lambda}_{b,j}$ in (5.9), $\lambda_{s,j}$, $\lambda_{r,j}$ and the thresholds β_Q, β_U , and β_L.

Figure 5.5 illustrates the impact of the upper and lower thresholds on MAP (i.e., β_U and β_L) for different collections at the concept layer $i = 2$. Because of the dependency between β_U and β_L in the approximate solution to the optimization problems, β_U and β_L are obtained iteratively one after the other by holding the other parameter fixed to a value obtained in the previous iteration. When the value of the upper threshold is less than the optimum, more non-useful concepts are added to the candidate list of expansion concepts. When the value of the upper threshold is greater than the optimum, some useful concepts may not be selected as expansion concepts. When the value of the lower threshold is less than the optimum, the selection process may terminate earlier, and a number of useful concepts may not be examined at all. When the value of the lower threshold is greater than the optimum, the proposed method will evaluate more concepts in total, which is against its main objective. Overall, although the upper and lower thresholds are dependent on each other, the upper threshold has the main effect on the accuracy of selected concepts, while the lower threshold has the main effect on the number of examined concepts.

Comparison of Methods

Table 5.5 provides comparison of performance of the proposed method with the baselines described in Section 5.3. As follows from this table, the best performing baseline is Method A, which is the most similar to the proposed method, since Method A and the proposed method both minimize the number of examined concepts. This can potentially reduce the effect of topic drift and results in superior performance of these methods.

The outermost concept layer, in which a method is able to identify the concepts that can increase the precision of retrieval results in another interesting criterion for method comparison.

A method that is able to identify effective expansion concepts in the outer concept layers is more robust, since these layers include higher number of noisy concepts. As follows from Table 5.5, the average outermost layer across different collections (rounded to the nearest integer), in which the baselines A-D and the proposed method were able to identify effective expansion concepts is 3, 3, 2, 2 and 4, respectively. Therefore, it can be concluded that the proposed method and the methods that have multiple thresholds tend to perform better than the methods that have a single threshold. The other conclusion that can be made from this table is that the average outermost layer across different collections (rounded to the nearest integer), in which the 4

Table 5.5. Comparison of retrieval performance of the proposed method with the baselines in terms of MAP for different number of examined concept layers.

Col.	Method	Concept Layers			
		1st	2nd	3rd	4th
TREC7–8	Method D-HAL	0.2220	0.2239	0.2155	0.2120
	Method D-CNet [48]	0.2205	0.2245	0.2214	0.2183
	Method C-HAL	0.2152	0.2227	0.2185	0.2133
	Method C-CNet	0.2182	0.2265	0.2225	0.2218
	Method B-HAL	0.2207	0.2171	0.2266	0.2236
	Method B-CNet	0.2188	0.2294	0.2255	0.2294
	Method A-HAL	0.2172	0.2251	0.2290	0.2282
	Method A-CNet	0.2183	0.2290	0.2329	**0.2335**
	Proposed-HAL	0.2249	0.2348	0.2418	**0.2457**
	Proposed-CNet	0.2222	0.2377	0.2449	**0.2484**
	SDM	0.2124	—	—	—
ROBUST04	Method D-HAL	0.2660	0.2644	0.2569	0.2554
	Method D-CNet [48]	0.2640	0.2651	0.2568	0.2555
	Method C-HAL	0.2675	0.2655	0.2608	0.2516
	Method C-CNet	0.2637	0.2628	0.2683	0.2695
	Method B-HAL	0.2684	0.2718	0.2598	0.2535
	Method B-CNet	0.2616	0.2710	0.2665	0.2675
	Method A-HAL	0.2614	0.2758	0.2757	0.2764
	Method A-CNet	0.2689	0.2732	**0.2851**	0.2793
	Proposed-HAL	0.2721	0.2786	0.2865	**0.2898**
	Proposed-CNet	0.2748	0.2814	0.2889	**0.2963**
	SDM	0.2359	—	—	—
GOV	Method D-HAL	0.2337	0.2428	0.2355	0.2319
	Method D-CNet [48]	0.2348	0.2396	0.2355	0.2382
	Method C-HAL	0.2404	0.2406	0.2459	0.2322
	Method C-CNet	0.2416	0.2451	0.2378	0.2379
	Method B-HAL	0.2359	0.2466	0.2418	0.2397
	Method B-CNet	0.2420	0.2452	0.2484	0.2421
	Method A-HAL	0.2434	0.2442	0.2491	0.2420
	Method A-CNet	0.2365	0.2455	**0.2524**	0.2422
	Proposed-HAL	0.2455	0.2429	0.2570	**0.2578**
	Proposed-CNet	0.2449	0.2514	0.2575	**0.2591**
	SDM	0.2184	—	—	—

Table 5.6. Comparison of retrieval performance of the proposed method with the baselines. * and † indicate statistically significant improvement in terms of MAP and P@20 according to Wilcoxon signed rank test over SDM/LCE with $p < 0.05$ and $p < 0.1$, respectively. Percentage differences in retrieval performance of Method A relative to SDM/LCE as well as the proposed method relative to SDM/LCE and Method A are shown in parentheses.

Without PRF							
Collection	**Evaluation Metric**	**QL**	**SDM**	**Method A HAL**	**Method A CNet**	**Proposed HAL**	**Proposed CNet**
TREC7–8	MAP	0.1982	0.2124	0.2282∗† (7.44%)	0.2335∗† (9.93%)	0.2457∗† (15.68%/7.67%)	0.2484∗† 16.95%/6.38%
	P@20	0.3540	0.3765	0.3762 (-0.08%)	0.3783 (0.48%)	0.3785∗ (0.53%/0.61%)	0.3796∗ 0.82%/0.34%
ROBUST04	MAP	0.2359	0.2510	0.2764∗† (10.12%)	0.2851∗† (13.59%)	0.2898∗† (15.46%/4.85%)	0.2963∗† 18.05%/3.93%
	P@20	0.3339	0.3667	0.3679 (0.33%)	0.3773∗† (2.89%)	0.3802∗† (3.68%/3.34%)	0.3795∗† 3.49%/0.58%
GOV	MAP	0.2184	0.2333	0.2491∗ (6.77%)	0.2524∗† (8.19%)	0.2578∗† (10.5%/3.49%)	0.2591∗† 11.06%/2.65%
	P@20	0.0389	0.0451	0.0476 (5.54%)	0.0493∗ (9.31%)	0.0558∗† (23.73%/17.23%)	0.0552∗† 22.39%/11.97%
With PRF							
Collection	**Evaluation Metric**	**RM**	**LCE**	**Method A* HAL**	**Method A* CNet**	**Proposed* HAL**	**Proposed* CNet**
TREC7–8	MAP	0.2151	0.2423	0.2503∗ (3.3%)	0.2558∗† (5.57%)	0.2642∗† (9.04%/5.55%)	0.2672∗† 10.28%/4.46%
	P@20	0.3641	0.3836	0.3883 (1.23%)	0.3927∗ (2.37%)	0.3934∗† (2.55%/1.31%)	0.4035∗† 5.19%/2.75%
ROBUST04	MAP	0.2683	0.2826	0.2935∗ (3.86%)	0.2979∗ (5.41%)	0.3034∗† (7.36%/3.37%)	0.3053∗† 8.03%/2.48%
	P@20	0.3561	0.3785	0.3826∗ (1.08%)	0.3834∗ (1.29%)	0.3893∗† (2.85%/1.75%)	0.3965∗† 4.76%/3.42%
GOV	MAP	0.2403	0.2678	0.2693 (0.56%)	0.2730∗ (1.94%)	0.2793∗† (4.29%/3.71%)	0.2811∗† 4.97%/2.97%
	P@20	0.0483	0.0566	0.0583 (3.00%)	0.0617∗ (9.01%)	0.0706∗ (24.73%/21.1%)	0.0720∗† 27.21%/16.69%

baselines and the proposed method were able to discover effective concepts, are 2 and 3 for the collection- and ConceptNet-based concept graphs, respectively. Overall, it can be also seen that the methods using ConceptNet-based concept graph (CNet) obtain higher MAP than the methods using collection-based concept graphs automatically constructed using HAL (HAL).

In Table 5.6, the performance of the proposed method is compared with QL, RM, SDM, LCE and the best performing methods in Table 5.5 that use collection- and ConceptNet-based concept graphs. As opposed to the upper part of Table 5.6, all the methods in its lower part also use unigram concepts from the top retrieved documents for query expansion, in addition to the

concepts from the concept graphs. The same collection- and ConceptNet-based concept graphs were used to obtain the results in the lower and upper parts of Table 5.6. The weights of the PRF unigram concepts were obtained using the RM model [51].

Several conclusions can be made from Table 5.6. First, Method A provides significant improvement over QL and SDM when the concept graph is generated by ConceptNet, while the proposed method has significant improvements over the baselines QL and SDM whether the concept graph is generated by HAL or ConceptNet. Second, Method A provides a significant improvement over SDM in the 5 cases, when it does not incorporate PRF concepts, however it provides a significant improvement over LCE only in one of the cases when it uses PRF concepts. Although the proposed method provides a smaller improvement over LCE, when it uses PRF concepts, than over SDM, when it does not use PRF concepts, the improvements that are achieved in these two cases are significant. Finally, although the parameters are estimated with the goal of maximizing MAP, the proposed method demonstrates significant improvement over the baselines (QE and SDM) also in terms of P@20.

CHAPTER 6 A BAYESIAN APPROACH TO UTILIZE KNOWLEDGE BASES IN MEDICAL INFORMATION RETRIEVAL

6.1. Introduction

IR methods for CDS have been the focus of several recent studies and evaluation campaigns. Specifically, the CDS track at the 2014–2016 Text Retrieval Conference (TREC) [95, 91, 89] sought to evaluate systems that provide evidence-based information in the form of full-text articles from the open access subset of PubMed Central to clinicians in response to medical case descriptions or admission notes as queries. The key challenges faced by these systems are the verbosity of queries, which include a complete account of patient visits, including details such as their vital signs and prescribed medications (e.g., queries in the CDS track of the 2016 TREC consist of the note, description, and summary fields with averages of 237, 120 and 33 terms, respectively); and vocabulary mismatch, which occurs when a query uses related concepts or different words to refer to the same concept in the relevant documents. To address these challenges, recently proposed systems [13, 12, 102] utilize techniques such as query interpretation, which involves locating clinical concepts using biomedical information extraction tools such as MetaMap [5], query expansion, which enriches the query with additional new terms, and query reduction, which removes terms with lower importance from the query [83, 27, 13, 12, 99].

Query expansion is one of the most effective techniques in boosting the retrieval performance in IR systems for CDS [83]. To enrich the query and alleviate the vocabulary mismatch problem, the query expansion approaches either use knowledge bases (such as the Unified Medical Language System (UMLS) and Medical Subject Headings (MeSH)), a collection of textual documents (such

as medical literature and electronic medical records) or a combination of them. One of the popular methods in the former approach is to obtain the expansion concepts through a pseudo relevance feedback (PRF) approach, i.e., by extracting concepts from top-ranked documents. Knowledge-based query expansion methods are useful when the knowledge base provides related concepts that also appear in the relevant documents. On the other hand, PRF-based query expansion approaches can improve the quality of the retrieval system if the initial list of retrieved documents is relevant enough to the query.

Table 6.1: An example of a query from the 2017 TREC precision medicine track [90].

Disease	Gastric cancer
Gene mutations	(PIK3CA, E545K)
Age	54
Gender	Male
Other	Depression

As in the general case of IR systems for CDS, the goal of IR systems for CDS in PM is to help healthcare providers find documents that are relevant to a patient case in an archive of biomedical articles. For example, a clinician may pose a query, such as that described in Table 6.1 that includes information about the cancer type, patient age, gender and other factors regarding the patient case, such as gene mutations. In general, queries posed to IR systems for CDS in PM have three distinct properties. First, these queries are significantly shorter than medical case descriptions. Therefore, the proposed method is focused on effective query expansion rather than on information extraction and concept weighting. Second, these queries are structured with the fields of queries of differing importance. Third, these queries contain both textual and non-textual information. Specifically, they typically include genetic variant data (e.g., mutations in

patient genes characterized by the gene name, such as “PIK3CA”, and amino acid (AA) position codes within the mutated gene, such as “E545K2”). Genetic variants play an important role in personalizing treatment because they can cause complex diseases, such as cancers, that share a similar set of symptoms to respond differently to the same treatment [8]. Therefore, the proposed method is focused on effectively incorporating gene mutation information into biomedical article retrieval.

Medical literature, such as PubMed, and medical data, such as UMLS, are two critical resources in the CDS systems. Bayesian networks provide a framework to utilize these two resources in the decision-making process. For example, Antal et al. [4] proposed to leverage medical literature to capture the prior belief in learning the dependency of two medical entities. In [4], the given data are considered as evidence to update the prior belief. In this work, we also learn a Bayesian network by incorporating medical literature and data from knowledge bases. In contrast to [4], in the IR problem we tackle, the prior knowledge regarding dependency of medical entities is provided by the medical knowledge bases, and the prior belief regarding the dependency of entities is updated given the query and its collection of medical literature. Besides, our Bayesian network is designed to facilitate retrieving medical articles that are relevant to a patient case under the PM paradigm.

To improve the accuracy of IR for CDS in the PM paradigm, we propose Bayesian Precision Medicine (BPM), which is a Bayesian approach for query expansion that utilizes information from knowledge bases as well as given queries. The focus of this work is to leverage relationships between mutated genes and candidate expansion concepts provided in the knowledge bases to perform query expansion. Because each mutated gene is often implicated in a variety of diseases

and can affect various tissues depending on each patient case, a naive automatic query expansion approach can deviate the topic (or aboutness) of the query away from the patient case. This problem, which is often called topic drift [62], makes many regular IR methods, such as the relevance feedback model [52], ineffective for this task [113]. We tackle this problem through our Bayesian approach and by utilizing a collection of medical documents, i.e., PubMed, and a genomic knowledge base, i.e., Catalog of Somatic Mutations in Cancer (COSMIC) [34].

BPM leverages the mentioned knowledge bases to compute a prior probability that a candidate concept for query expansion is related to a mutated gene mentioned in the query. Then, by using this prior probability and information provided in the query, it computes a posterior probability of a candidate concept being related to a given query. The main challenge that we addressed in this medical IR task is the limitation on the size of the training data that was obtained by medical experts for a limited number of queries3. In this IR task, to address the vocabulary mismatch problem, features from multiple resources (knowledge bases and collection of medical literature) are required which makes the traditional query expansion methods to perform poorly due to their need for a large number of training data. To tackle this problem, we introduce a number of assumptions in our method to simplify its training process.

It is worthwhile to highlight the main contributions of this proposed method. (1) We are the first to introduce a Bayesian approach for expanding medical queries in a PM paradigm. (2) We provide a comprehensive analysis of our method under different scenarios of extracting concepts from the collection of biomedical articles and knowledge bases and under different configurations of our model.

6.1. Method

BPM selects a list of concepts for query expansion via a Bayesian approach by measuring the relatedness of these concepts to the query. BPM leverages information from knowledge bases as well as given queries combining textual and genomic information. As depicted in Figure 6.1, BPM executes the following steps for query expansion:

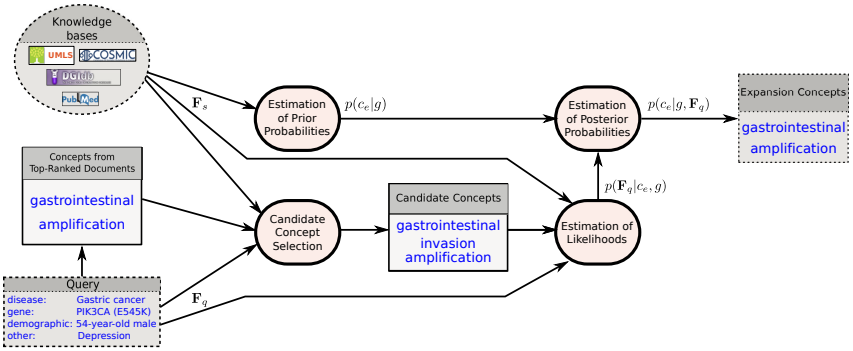


Figure 6.1: The architecture of our Bayesian method (BPM) that leverages multiple knowledge bases to measure the relatedness of candidate expansion concepts to the given query in a precision medicine paradigm.

1. BPM generates a list of candidate concepts for query expansion (such as “PIK3C4”, “gastrointestinal”, ⋯) by using UMLS and the Drug-Gene Interaction Database (DGIdb) (see Section 6.2).
2. Using COSMIC and PubMed, BPM estimates the prior probability of relatedness of a candidate concept (such as “PIK3C” or “gastrointestinal”) to a mutated gene mentioned in the query (such as “PIK3CA”) (see Section 6.2).
3. Using PubMed, BPM estimates the likelihood of having a patient case described by features in the query (such as those in Table 6.1) given a mutated gene and a candidate concept (see Section 6.2).

4. Using prior probabilities and previously estimated likelihoods, BPM estimates posterior probabilities to determine whether to accept a candidate expansion concept (see Section 6.2).

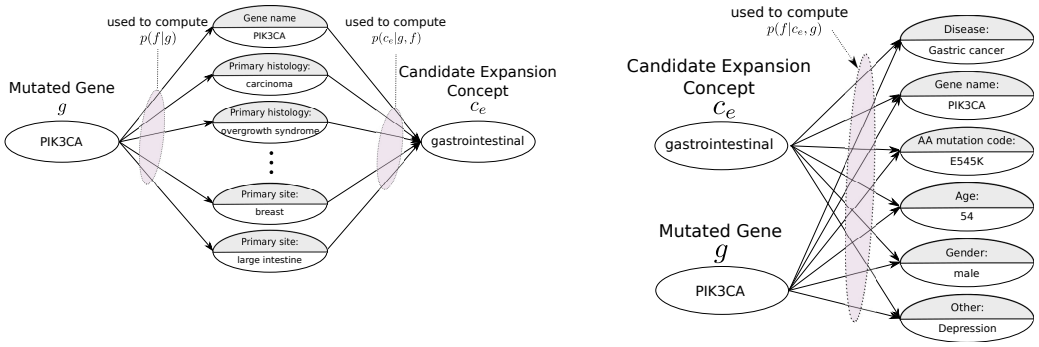
Candidate Expansion Concepts:

Given each patient case, BPM generates a list of candidate expansion concepts that likely fills the vocabulary gap between a query and its relevant biomedical articles and ultimately improves the retrieval accuracy. As shown in Figure 6.3, BPM utilizes the following three sources to select candidate concepts for query expansion:

1. UMLS table of relationships (MRREL): BPM selects concepts that are related to diseases or gene names in the query according to MRREL.
2. Top-ranked documents from PubMed: BPM selects concepts that appear in top-ranked documents for the original query. In the experiments of this work, we extract 40 top-ranked concepts according to the relevance model [52] from 25 top-ranked documents.
3. DGIdb [38]: BPM selects the names of drugs that have interactions with mutated genes according to DGIdb.

The Bayesian Approach

Each given query, which describes a patient case, represented by the set of features F_p. BPM estimates the probability of relatedness of a candidate expansion concept c_e to a query (i.e., $p(c_e|F_p)$). As described in Section 6.2 and shown in Figure 6.2(b), the set of features F_p, which BPM directly extracts from the query fields, contains six feature types (“Disease”, “Gene name”, “AA mutation”, “Age”, “Gender” and “Other”). We denote the name of the mutated gene



(a) An illustration of features extracted from the COSMIC knowledge base used in (6.7) to compute the prior probability of relatedness of a candidate expansion concept to a mutated gene.

(b) An illustration of features extracted from the query used in (6.11) to compute the likelihood of the query given a candidate expansion concept and a mutated gene.

Figure 6.2: An illustration of the Bayesian networks used to incorporate information from the COSMIC knowledge base (i.e., "Gene name", "AA mutation code", "Primary site", and "Primary histology") and information from the query (i.e., "Disease", "Gene name", "AA Mutation code", "Age", "Gender", and "Other") to compute the prior probability in (6.7) and the likelihood in (6.11).

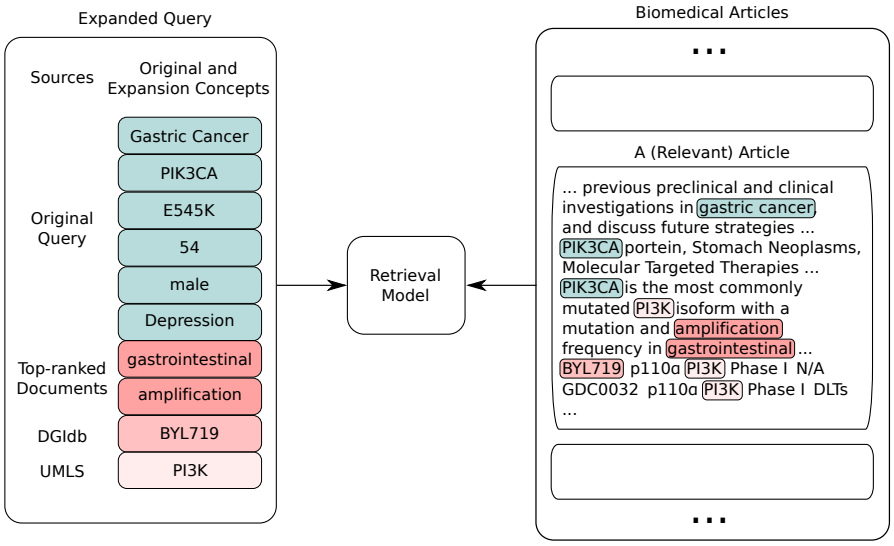


Figure 6.3: An illustration of the process to retrieve biomedical articles for the example query in Table 6.1 expanded by using the sources described in Section 6.2. This figure shows that by expanding the query with concepts from different sources, BPM fills the vocabulary gap between a query and its relevant document in the collection. In this figure, "gastrointestinal" and "amplification" are concepts from top-ranked documents, BYL719 (Phosphatidylinositol 3-Kinase α-Selective Inhibition With Alpelisib) is from DGIdb, and PI3K (Phosphoinositide 3-kinase) is from UMLS table of relationships.

mentioned in the query by a code, such as “PIK3CA”, by g. As we will show later, the mutated gene g is often the most important feature in the set of features extracted from the query (F_p). Therefore, we separate g from the rest of features in the set F_p and write the posterior probability $p(c_e|F_p)$ as follows:

$$p(c_e|F_p) = p(c_e|g, F_q) \quad (6.1)$$

where $F_p = \{g\} \cup F_q$. By following a Bayesian approach, BPM computes $p(c_e|F_q)$ by first learning the prior probability of the candidate expansion concept c_e being related to the mutated gene g mentioned in the query ($p(c_e|g)$) thorough incorporating features from the COSMIC knowledge base (F_s). Next, by incorporating more information from the query, BPM assesses the

Table 6.2: Table of Notations.

Notation	Definition
c_e	candidate expansion concept, such as “gastrointestinal”
F_s	set of features extracted from the COSMIC knowledge base
F_p	set of features extracted from the query that describes a patient case
F_q	set of features extracted from the query excluding the mutated gene name
g	mutated gene name, such as “PIK3CA”
$p(c_e\|f)$	probability of concept c_e being related to feature f
$p(f\|g)$	probability of an association of feature f with mutated mutated gene g
$p(F_q\|c_e, g)$	likelihood of having a patient case represented by F_q conditioned on concept c_e and mutated gene g
$p(c_e\|g)$	prior probability of concept c_e being related to mutated gene g
$p(c_e\|g, F_q)$	posterior probability of concept c_e being related to mutated gene g and features in F_q extracted from query

relatedness of the candidate expansion concept c_e to the query and decides whether the candidate concept c_e should be selected as an expansion concept for a given query.

By following a similar formulation as in [96], we use Bayes' rule to rank the candidate expansion concept c_e based on its relatedness to the query as follows:

$$posterior = p(c_e|g, F_q) = \frac{p(c_e, F_q|g)\, p(F_q|g)}{p(Fq|g)}$$

which can be rewritten as:

$$posterior = \frac{p(c_e|g)p(F_q|ce,g)}{p(Fq|g)} = \frac{prior \times likelihood}{constant} \quad (6.2)$$

Because we compute $p(c_e|g, F_q)$ to rank the candidate concepts and the denominator in the above equation is a constant of the candidate expansion concept c_e, we can write:

$$p(c_e|g, F_q) \propto p(c_e|g)p(F_q|c_e, g) \quad (6.3)$$

This Bayesian approach is demonstrated in Figure 6.2. As shown in Figure 6.2, in the first step (described in Section 6.2), by utilizing only the knowledge bases, we estimate the prior probability of relatedness of each candidate expansion concept to the mutated gene mentioned in the query. As explained in sections 6.2 and 6.2, at the next step, by utilizing the information provided in the query, we assess the relatedness of each candidate expansion concept to the query. Then, as explained in Section 6.2, we select the expansion concepts according to their relatedness to the query.

Features

To estimate the relatedness of a candidate expansion concept to a query, BPM leverages two sets of features, $\mathbf{F}_s$ and $\mathbf{F}_q$, extracted from the COSMIC knowledge base and a query, respectively. As illustrated in Figure 6.2(a), the sets of features Fs, which are extracted from the

table of *COSMIC mutation data*[6], contain a list of approximately 6 million records of patients with a type of cancer that may have been caused by one of approximately 30,000 gene mutations. $\mathbf{F}_s$ includes features of type "Gene name", "AA mutation code", "Primary site", and "Primary histology" that are extracted from COSMIC. Table 6.3 presents an illustration of data extracted from COSMIC for the mutated gene "PIK3CA", which BPM utilizes to extract the features in $\mathbf{F}_s$ given PIK3CA as the mutated gene. For example, given "large intestine" as the value of feature with type "Primary site", BPM uses this table by counting the number of samples (patient records) in this table that have a mutation in gene PIK3CA and have "large intestine" as the primary site (tissue) affected by this mutated gene. This will be described in more detail in the next section. As illustrated in Figure 6.2(b), the set of features extracted from a query ($\mathbf{F}_q$) include "Disease", "Gene name", "AA mutation code", "Age", "Gender", and "Other". BPM directly extracts these features from the fields of the queries illustrated in Table 6.1. BPM extracts the set of features "Gene name" and "AA mutation code" from the query field "gene", while it extracts the features "Age" and "Gender" from the query field "demographic". The features "Disease" and "Other" are described by the query fields "disease" and "other", respectively. The "Other" feature contains factors that are related to the patient case, such as symptoms, secondary diseases, and surgical procedures.

BPM also incorporates term frequency and co-occurrence features measured from the collection of medical articles in PubMed in computing the prior probability ($p(c_e|g)$) and likelihood ($p(\ F_q|c_e, g)$). We used a January 2017 snapshot of 26.8 million PubMed abstracts[7]. As

[6] Publicly available at https://cancer.sanger.ac.uk/cosmic/download\#download-3. We used COS-MIC v84, released 13-FEB-18.

[7] Publicly available at https://bionlp.nlm.nih.gov/trec2017precisionmedicine/

described in Section 6.2, BPM uses UMLS, PubMed, and DGIdb as sources to generate a list of candidate expansion concepts, while, as described in this section, BPM uses PubMed and COSMIC as sources to measure the relatedness of these concepts to a query.

Computing Prior Probabilities Given Knowledge Bases

Given the PubMed and COSMIC knowledge bases, BPM computes the prior probability of the relatedness of the concept c_e to the mutated gene g (i.e., $p(c_e|g)$). In computing the prior probability $p(c_e|g)$, BPM incorporates COSMIC to provide information regarding a mutated gene, such as its primary sites (tissues) and histologies. To do so, BPM extracts the set of features Fs from the COSMIC knowledge base. In this work, for simplicity, we only consider four feature types "Gene name", "AA mutation code", "Primary site", and "Primary histology".

BPM computes the prior probability $p(c_e|g)$ by averaging over all possible assignments to all features in F_s, which are extracted from the COSMIC knowledge base. In other words,

$$p(c_e|g) = \sum_{f_1,\ldots,f_n} p(c_e, F_1 = f_1, \ldots, F_n = f_n|g) \quad (6.4)$$

To compute $p(c_e|g)$, we need to estimate an exponential number of probabilities which is not feasible in practice when the number of features (n) is large. Therefore, following similar approach as [29, 4], we assume an Independence between features in the Bayesian network to be able to compute $p(c_e|g)$ with a linear number of probabilities. Therefore, we estimate $p(c_e|g)$ as

$$p(c_e|g) \approx \sum_f p(F = f|g)p(c_e|g, F = f) \quad (6.5)$$

where f is a value of feature $F \in \mathbf{F}_s$. $p(F|g)$ determine the importance of feature F in computing $p(c_e|g)$ and is dependent on the mutated gene g mentioned in the query. In other words, $p(c_e|g)$ accounts for the differences in the importance of different features (i.e., "Gene

name", "AA mutation code", "Primary site", and "Primary histology") in computing the prior probabilities.

As an example, if we consider the feature F to have only one type, e.g., "Primary site" (see Table 6.3), the sum in the above equation is over all possible primary sites affected by a mutated gene g according to the COSMIC knowledge base. In the example illustrated in Table 6.3, the feature "Primary site" has five distinct values.

The probabilities $p(c_e|g,f)$ and $p(f|g)$ in the above equation can be interpreted as follows:

1. $p(f|g)$: the probability of association of feature f with the mutated gene g, and

2. $p(ce|g,f)$: the probability of the concept $c_e$being related to the mutated gene g and feature f.

Table 6.3: An example of data extracted from COSMIC for the gene "PIK3CA". Only 12 of 13,120 columns corresponding to gene g = "PIK3CA" are shown in this figure. According to the COSMIC database, "breast" and "large intestine" are among the most affected primary sites when the gene "PIK3CA" has a mutation.

Gene name	AA mutation code	Primary site	Site subtype 1	Primary histology	Histology subtype 1	Age
PIK3CA	p.E545K	large intestine	—	carcinoma	adenocarcinoma	—
PIK3CA	p.M1043T	stomach	—	carcinoma	adenocarcinoma	—
PIK3CA	p.E545K	large intestine	—	carcinoma	adenocarcinoma	—
PIK3CA	p.E542K	stomach	—	carcinoma	intestinal adenocarcinoma	74
PIK3CA	p.H1047R	breast	—	carcinoma	—	—
PIK3CA	p.H1047R	breast	—	carcinoma	ductal carcinoma	—
PIK3CA	p.E545K	breast	—	carcinoma	ductal carcinoma	—
PIK3CA	p.E545K	breast	—	carcinoma	—	—
PIK3CA	p.E545K	lung	—	carcinoma	bronchioloalveolar adenocarcinoma	67
PIK3CA	p.E545K	large intestine	right	carcinoma	adenocarcinoma	73
PIK3CA	p.H1047R	large intestine	colon	carcinoma	adenocarcinoma	—
PIK3CA	p.E545K	soft tissue	fat	Overgrowth syndrome	CLOVES syndrome	14

We estimate the probability $p(f|g)$ from the COSMIC mutation data. Table 6.3 presents an example of data extracted for the gene PIK3CA from the COSMIC database. The probability $p(f|g)$ is estimated from COSMIC as follows:

$$p(f|g) = \frac{p(f,g)}{p(g)} \approx \frac{N_{f,g}}{N_g}, \tag{6.8}$$

where N_g is the number of cancer patients who have mutations in gene g and $N_{f,g}$ is the number of cancer patients who have mutations in gene g and are associated with feature f according to the COSMIC database.

The probability $p(c_e|f,g)$ is estimated by using the collection of medical articles in PubMed as the knowledge base. We represent both concept c_e and the values of feature f by ngrams, such as “high blood pressure” and “colon cancer”. BPM uses PubMed to measure the degree of the semantic relationship of concept c_e to mutated gene g and feature f. To do so, we find documents that contain ngram representations of feature f and mutated gene g (i.e., $R_{f,g}$), and among these documents, we find the portion that also contains the $n$gram representation of concept c_e, shown by $R_{c_e,f,g}$. Therefore, we estimate the probability $p(c_e|f,g)$ as follows:

$$p(c_e|f,g) = \frac{p(c_e,f,g)}{p(f,g)} \approx \frac{|R_{c_e,f,g}|}{|R_{f,g}|}. \quad (6.9)$$

where $|R_{c_e,f,g}|$ and $|R_{f,g}|$ are the number of documents in the sets $R_{c_e,f,g}$ and $R_{f,g}$, and $|R_{c_e,f,g}| \leq |R_{f,g}|$, respectively, since $R_{c_e,f,g} \subseteq R_{f,g}$. To avoid a zero-frequency problem, we use the following smoothing method [123]:

$$p(c_e|f,g) \approx \frac{\beta+|R_{c_e,f,g}|}{\beta N + |R_{f,g}|} \quad (6.10)$$

where N is the number of documents in the collection and β is a constant that we consider to equal 10^{-6}.

Example. As an example, we assume BPM computes the prior probability of the concepts c_e= “PI3K” and c_e = “exon” to the gene g = “PIK3CA” by using COSMIC and PubMed as the knowledge bases. For simplicity, in this example, we take a single feature (“Primary site”) as the only feature

extracted from the COSMIC database and use the samples shown in Table 6.3 to compute the prior probability $p(c_e|g)$. By using the maximum likelihood (ML) approach in (6.8), we obtain

$$p(c_e = \text{“PI3K”}|g = \text{“PIK3CA”}) = 0.239$$

and

$$p(c_e = \text{“exon”}|g = \text{“PIK3CA”}) = 0.087$$

Table 6.4: An illustration of steps to compute the prior probability of the candidate expansion concept c_e being related to gene g. For the sake of illustration, the probability $p(f|g)$ is computed using only the sample data shown in Table 6.3. The values of $|R_{f,g}|$ and $|R_{c_e,f,g}|$ and as a result $p(c_e|f,g)$ are estimated by using the PubMed collection.

Gene g	Concept c_e	Feature Type F	Feature Value f	$p(f\|g)$	$\|\mathcal{R}_{f,g}\|$	$\|\mathcal{R}_{c_e,f,g}\|$	$-\log p(c_e\|f,g)$
PIK3CA	PI3K	Primary site	breast	4/12	994	443	0.93
PIK3CA	PI3K	Primary site	large intestine	4/12	8	1	2.16
PIK3CA	PI3K	Primary site	stomach	2/12	12	6	1.57
PIK3CA	PI3K	Primary site	lung	1/12	660	154	1.39
PIK3CA	PI3K	Primary site	soft tissue	1/12	33	11	1.78

(a)

Gene g	Concept c_e	Feature Type F	Feature Value f	$p(f\|g)$	$\|\mathcal{R}_{f,g}\|$	$\|\mathcal{R}_{c_e,f,g}\|$	$-\log p(c_e\|f,g)$
PIK3CA	exon	Primary site	breast	4/12	994	167	1.81
PIK3CA	exon	Primary site	large intestine	4/12	8	0	17.36
PIK3CA	exon	Primary site	stomach	2/12	12	2	2.97
PIK3CA	exon	Primary site	lung	1/12	660	142	1.57
PIK3CA	exon	Primary site	soft tissue	1/12	33	5	2.48

(b)

We can observe from this example that the computed prior probability can distinguish a concept such as “PI3K”, which is more related to gene “PIK3CA” from a concept such as “exon”, which is related to all genes in general.

Computing Likelihoods Given Query and Knowledge Bases

Given a patient case in the form of a query, BPM assesses the relatedness of the candidate expansion concept c_e to the query. To do so, first, BPM computes $p(\mathbf{F}_q|c_e,g)$, which is the

likelihood of having a patient case represented by the query features $\mathbf{F}_q$ conditioned on the candidate expansion concept c_e and mutated gene g. For example, BPM may compute the likelihood of having a patient case described in Table 6.1 conditioned on having a mutation in gene "PIK3CA" and having the concept "large intestine" as a candidate expansion concept.

By following a similar approach in the previous section in simplifying the computation of prior probability, we can estimate the likelihood $p(\mathbf{F}_q|c_e, g)$ through a Naive Bayes conditional independence assumption [53] over features extracted from the query ($\mathbf{F}_q$). Based on this assumption, we can estimate $p(\mathbf{F}_q|c_e, g)$ from the probability of feature $f \in \mathbf{F}q$ being related to concept c_e and gene g as follows:

$$p(\mathbf{F}_q|c_e, g) = \prod_{f_1,\ldots,f_n} p(F_1 = f_1, \ldots, F_n = f_n|c_e, g)$$

$$\approx \prod_f p(F = f|c_e, g)$$

Since the features in $\mathbf{F}_q$ have significantly different importance, we estimate the likelihood in the above equation as

$$\log p(\mathbf{F}_q|c_e, g) \approx \sum_f v_f \log p(F = f|c_e, g) \quad (6.11)$$

where $0 < \omega_f < 1$ depends on the importance of feature type F in computing the likelihood $p(\mathbf{F}_q|c_e, g)$. We assume the importance of a feature type (ω_f) to be independent of the mutated gene (g) mentioned in a query and the candidate expansion concept (c_e). We obtain ω_f through the cross-validation process. As an example, if we consider f to be the feature "Other" extracted from the query, the second product in the above equation is over all the phrases listed in the query field "Other", and v_f is the weight of feature "Other" in computing the likelihood. The probability $p(f|c_e, g)$ is computed in a manner similar to that described in (6.9) as follows:

$$p(f|c_e, g) = \frac{p(f,c_e,g)}{p(c_e,g)} \approx \frac{|R_{f,c_e,g}|}{|Rc_e,g|} \quad (6.12)$$

To avoid a zero-frequency problem, we use the following smoothing method [123]:

$$p(f|c_e, g) \approx \frac{\beta+|R_{f,c_e,g}|}{\beta N+|R_{c_e,g}|} \quad (6.13)$$

where N is the number of documents in the collection and β is a constant that we consider to equal 10^{-6} in our experiments.

Example. As an example, we assume BPM aims to compute the likelihood $p(\mathbf{F}_q|c_e, g)$ for the set of features $\mathbf{F}q$ being extracted from the example query illustrated in Table 6.1 conditioned on the concept c_e ="gastrointestinal" and mutated gene g ="PIK3CA". This example is illustrated in Figure 6.2(b) for the case of extracting the following set of features from the example query:

$$F_q = \{\text{“}Disease\text{”} = \text{“}Gastric\ cancer\text{”}, \text{“}Gene\ name\text{”} = \text{“}PIK3CA\text{”}, \text{“}AA\ mutation\ code\text{”} = \text{“}E545K\text{”}, \text{“}Age\text{”} = 54, \text{“}Gender\text{”} = \text{“}Male\text{”}, \text{“}Other\text{”} = \text{“}Depression\text{”}\}$$

As Table 6.5 illustrates, if we provide the same weights for all features, we obtain log likelihoods as

$$\log(p(F_q|c_e = \text{“}gastrointestinal\text{”}, g = \text{“}PIK3CA\text{”})) = -4.92$$

and

$$\log(p(F_q|c_e = \text{“}express\text{”}, g = \text{“}PIK3CA\text{”})) = -5.68,$$

which reveals that when the candidate expansion concept is a general concept, such as “express”, it has a lower likelihood than the case of a candidate expansion concept that is more related to the original query, such as “gastrointestinal”. Later, we will show that by providing different weights for different features based on their importance, the computed likelihood and

consequently the computed posterior probability provide a better understanding of the relatedness of a candidate expansion concept to the query.

Computing Posterior Probabilities Given Query and Knowledge Bases

As the final step, by using (6.3), BPM computes the posterior probabilities for all candidate expansion concepts and ranks them accordingly. To achieve this goal, from the prior probabilities $p(c_e|g)$ computed in Section 6.2 and likelihoods $p(F_q|c_e, g)$ computed in Section 6.2, BPM uses the Bayes' rule to compute the posterior probability $p(c_e|g, F_q)$ for each candidate expansion concept c_e. In other words, BPM turns the prior belief (computed by using (6.7)) by incorporating evidence about the patient case (computed by using (6.11)) into a posterior belief (computed by using (6.3)).

Representing Concepts

The term dependencies play an important role in representing the original and expansion medical concepts in the retrieval models [27, 13]. Similarly, BPM uses an SDM to capture the term dependencies of concepts in its retrieval model. As described in the next section, in the retrieval model of BPM, the relatedness of a document to a query is computed from the relatedness of that document to the original and expansion concepts in the query. Given the concept "gastric cancer", BPM computes the relevance of a document to this concept by computing the following:

1. the numbers of times that the concept terms "gastric" and "cancer" appear in the given document,
2. the number of windows in the given document that contain "gastric" and "cancer" in the same order as that in the n-gram representation of the concept, and

3. the number of windows in the given document that contain the mentioned two terms in any order.

Table 6.5: An illustration of steps in our method to compute the relatedness of feature *f* extracted from the query to the candidate expansion concept c_e and gene g.

concept c_e	gene g	feature type F	feature value f	$\lvert\mathcal{R}_{c_e,g}\rvert$	$\lvert\mathcal{R}_{f,c_e,g}\rvert$	$-\log p(f = x_f \lvert c_e, g)$
"gastrointestinal"	"PIK3CA"	"Disease"	"Gastric cancer"	121	16	2.22
"gastrointestinal"	"PIK3CA'	"Gene name"	"PIK3CA"	121	121	0.02
"gastrointestinal"	"PIK3CA'	"AA mutation code"	"E545K"	121	3	3.89
"gastrointestinal"	"PIK3CA'	"Age"	"54"	121	18	2.10
"gastrointestinal"	"PIK3CA'	"Gender"	"Male"	121	12	2.51
"gastrointestinal"	"PIK3CA'	"Other"	"Depression"	121	0	18.81
"express"	"PIK3CA"	"Disease"	"Gastric cancer"	1374	57	3.20
"express"	"PIK3CA'	"Gene name"	"PIK3CA"	1374	1374	0.02
"express"	"PIK3CA'	"AA mutation code"	"E545K"	1374	58	3.18
"express"	"PIK3CA'	"Age"	"54"	1374	54	3.26
"express"	"PIK3CA'	"Gender"	"Male"	1374	23	4.10
"express"	"PIK3CA'	"Other"	"Depression"	1374	0	20.33

BPM normalizes these three values over the size of the document and obtains a weighted linear combination of them to compute the relevance of a given document to a query concept. BPM repeats this process for all of the original and expansion concepts in the query and computes the relevance of document D to the given patient case described in the query. We obtain the sizes of the mentioned windows via the cross-validation process. We formulate the ranking of collection documents given n-gram concepts in the next section.

Ranking Candidate Concepts and Collection Documents

BPM utilizes a retrieval system that is composed of two steps of

1. ranking candidate concepts to expand the query and

2. ranking collection documents given the expanded query.

In the first step, BPM selects candidate expansion concept c_e as the expansion concept if $p(c_e|g, F_q)$ computed by (6.2) goes above a threshold. We obtain this threshold through the cross-validation process.

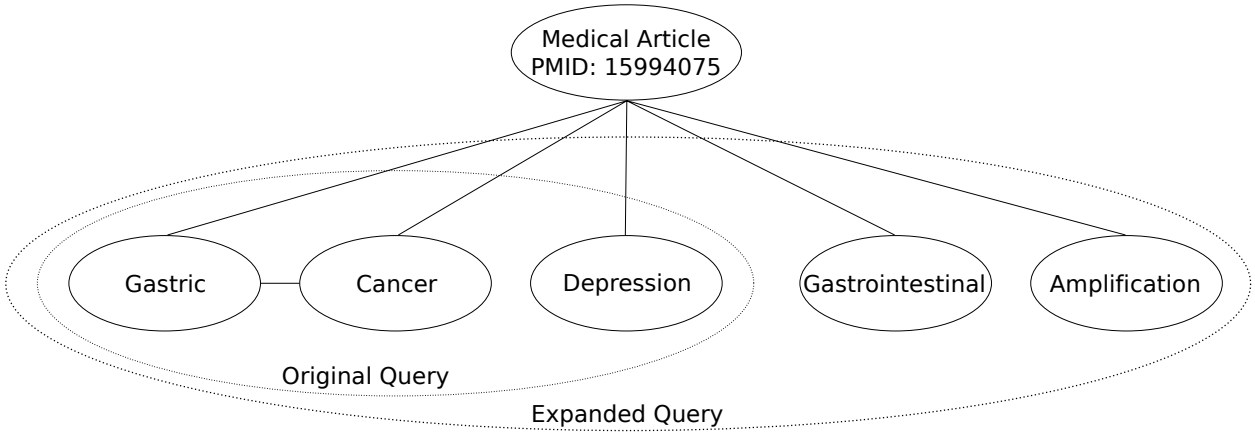


Figure 6.4: An illustration of the graphical representation of SDM for our query expansion method in a PM task. In this illustration, the query has two original concepts (Gastric cancer, and Depression) and two expansion concepts (Gastrointestinal and Amplification). This query is extracted from the query set of 2017 TREC-PM task, and the medical article with PMID (PubMed ID) 15994075 (https://www.ncbi.nlm.nih.gov/pubmed/15994075) is a relevant article for this query. Adding the expansion concepts to the query alleviates the vocabulary mismatch problem between the query and its relevant documents since they often appear in the relevant articles but do not exist in the original query.

As can be seen from the example illustrated in Figure 6.3 and the graphical representation in Figure 6.4, at the second step, BPM ranks the collection documents given the expanded query ($\tilde{Q}$) according to [16] as follows:

$$s(D, \widetilde{Q}) = s(\text{document}, \text{original query}) + \alpha_e s(\text{document}, \text{expansion concepts}) \quad (6.14)$$

where α_e depends of the importance of expansion concepts in comparison to the original query concepts in computing the relevance score $s(D, \widetilde{Q})$. This equation can be rewritten as:

$$s(D, \tilde{Q}) = \sum_{l \in Q} \omega_l s(D, Q_l) + \sum_{x \in X} \omega_x s(D, \tilde{Q}_x) \quad (6.15)$$

where Q_l is a field in the original query Q. As shown in Figure 6.3, the fields in the original query are disease name, mutated gene name, AA mutation code, age, gender, and other. In the above equation, Q, $\tilde{Q}_x$, and $\tilde{Q}$ contain original concepts from the query, expansion concepts extracted from the source x, and all expansion and original concepts, respectively. In the above equation,

X is the list of sources for the query expansion introduced in Section 6.2. We can observe from the above equation that different fields of the query and different expansion concept sources receive different weights (i.e., ω_l and ω_x) in computing the score.

By using SDM, illustrated in Figure 6.4, BPM computes the score $s(D, Q_l)$ (and similarly $s(D, \tilde{Q}_x)$) as follows:

$$s(D, \tilde{Q}_l) = \alpha_T \, s(\text{document, unigrams in } Q_l) + \alpha_U \, s(\text{document, ordered bigrams in } Q_l) + \alpha_O \, s(\text{document, unordered bigrams in } Q_l) \quad (6.16)$$

All of the three score functions in the left-hand side of (6.16) are computed from the collection of medical articles in PubMed. For example, if Q_l = "Gastric Cancer", then the first score function in (6.16) depends on the number of articles that have words "Gastric" or "Cancer", the second score function in (6.16) depends on the number of articles that have the word "Gastric" after the word "Cancer" in windows of limited size, and the third score function in (6.16) depends on the number of articles that have words "Gastric" and "Cancer" in any order in windows of limited size. The above equation can be rewritten as

$$s(D, Q_l) = \lambda_T \sum_{u \in Q_l} f_T(D, u) + \sum_{b \in Q_l} \left(\lambda_U \, f_U(D, b) + \lambda_O \, f_O(D, b) \right) \quad (6.17)$$

where u and b are a unigram and a bigram in Q_l. $f_T(D, u)$ is a feature function that determines the score of the collection document D given a unigram in the query. The feature functions $f_O(D, b)$ and $f_U(D, b)$ determine the score of the collection document D given the bigram b in the query with and without considering the order of the terms in the bigram.

The feature function $f_T(D,u)$ is computed by using Bayesian smoothing with Dirichlet priors [123] from the count of unigrams u in document D $(tf_{u,D})$ and in the collection (cf_u) as follows:

$$f_T(D,u) = \log\left(\frac{tf_{u,D}+\mu\frac{cf_u}{|C|}}{|D|+\mu}\right) \tag{6.18}$$

where $|C|$ is the number of terms in the collection (i.e., length of the collection), $|D|$ is the number of terms in document D (i.e., length of document D), and μ is a constant. BPM computes $s(D,\tilde{Q})$ and $s(D,\tilde{Q}_x)$ in a manner similar to that described above.

We define $tf_{\#od(b),D}$ and $tf_{\#uw(b),D}$ as the numbers of windows with sizes n_O and n_U in the document D that contain constituent terms of bigram b, respectively, in the same order as that in the concept's bigram representation and in any order. For example, for the concept "long intestine", $tf_{\#od(b),D}$ is the number of windows that contain the term "long" before the term "intestine", and $tf_{\#uw(b),D}$ is the number of windows that contain these two terms in any order. BPM computes $f_O(D,b)$ and $f_U(D,b)$ as follows:

$$f_O(D,b) = \log\left(\frac{tf_{\#od(b),D}+\mu\frac{cf_{\#od(b)}}{|C|}}{|D|+\mu}\right) \tag{6.19}$$

and

$$f_U(D,b) = \log\left(\frac{tf_{\#uw(b),D}+\mu\frac{cf_{\#uw(b)}}{|C|}}{|D|+\mu}\right) \tag{6.20}$$

where $cf_{\#od(b)} = \sum_D tf_{\#od(b),D}$ and $cf_{\#uw(b)} = \sum_D tf_{\#uw(b),D}$ are the numbers of windows in all documents in the collection that contain bigram b with and without considering whether the order of terms is the same as that in the concept's representation, respectively. We obtain the window sizes n_O and n_U and the constant μ in the above equation via the cross-validation

process. For ngram concepts with $n > 2$, BPM divides them into multiple bigrams. For example, BPM represents the concept "malignant gastric ulcer" in the form of the two bigrams "malignant gastric" and "gastric ulcer".

Consideration of Mutated Gene vs Gene Mutation in Computing the Prior Probabilities

As an alternative to our solution described above, which is to compute the prior probability of a candidate expansion concept c_e being related to mutated gene g in the query (e.g., $p(c_e = \text{"gastrointestinal"}|g = \text{"PIK3CA"})$), the prior probability of a candidate expansion concept being related to gene mutation m can be computed (e.g., $p(c_e = \text{"gastrointestinal"}|m = \text{"PIK3CA (E545K)"})$. Here, gene mutation m is represented by a gene name and an AA mutation code (e.g., "PIK3CA (E545K)"). However, due to the sparsity of these ngrams in PubMed, the latter solution requires a collection much larger than PubMed to compute $p(c_e|m)$ and $p(c_e|m, F_q)$. For example, in a collection of 26 million articles in PubMed, there are only 192 documents that contain both terms "PIK3CA" and "E545K" in any order, and there are only 3 documents that contain the terms "PIK3CA", "E545K" and "gastrointestinal" in any order. Therefore, instead of m, we propose to consider the mutated gene name g in computing the prior probability. We will discuss this solution in more detail in Section 6.3.

6.3. Experiments

Dataset and Implementation Details

Our training data consist of 18,729 unique medical articles whose relevance to at least one of the 30 available queries was judged by experts in the field [90]. These medical articles were obtained from PubMed and proceedings of the American Association for Cancer Research

(AACR)[8] and the American Society of Clinical Oncology (ASCO)[9]. As shown in Figure 6.5, of 22,642 total judgments, the articles were judged as "Not Relevant" in most cases (82.89% of the judgments), and "Definitely Relevant" and "Partially Relevant" in only 8.93% and 8.18% cases, respectively. For more details regarding the relevance assessment steps taken to gather the relevance judgments, please see [90]. Other than a January 2017 snapshot of 26 million PubMed documents, 70,025 abstracts from the AACR and ASCO proceedings were adopted as collections to evaluate BPM[10]. These medical articles, the queries, and their relevance judgments for the PM task were provided by the 2017 TREC-PM track and are publicly available[11]. The 2017 TREC-PM track contains a task of retrieving clinical trials (from ClinicalTrials.gov) which is beyond the scope of this work. There are 30 patient cases described in the form of queries[12], and their corresponding lists of relevance judgments[13] with their full annotations (including disease, gene, etc.)[14] were provided by the National Institute of Standards and Technology (NIST).

We use the Indri search engine [105] to index the medical articles in the collection and to run the queries. We used MetaMap to map the phrases in the query to their UMLS concept IDs. We only index the following fields of the PubMed articles because these fields often contain the most important information about the article:

1. Article Title,

2. Abstract,

[8] http://www.aacr.org/
[9] https://www.asco.org/
[10] https://bionlp.nlm.nih.gov/trec2017precisionmedicine/
[11] http://www.trec-cds.org/2017.html
[12] http://www.trec-cds.org/topics2017.xml
[13] http://www.trec-cds.org/qrels-treceval-abstracts.2017.txt
[14] https://drive.google.com/open?id=1IH4dL4OKG7bv57K8DreOeSAfJgkgC4sd

3. MeSH Headings List, and

4. Chemical List.

These fields of PubMed articles were extracted from the XML files provided by the 2017 TREC-PM track. The MeSH Headings List contains MeSH terms that are most related to each article. The Chemical List contains the MeSH terms of the chemical compounds described in each article. In the 2017 TREC-PM track, articles from AACR and ASCO proceedings are represented by only their titles and abstracts, and we index only these two fields in our work. We use a three-fold cross-validation strategy to tune the hyper-parameters of BPM and the baselines. To compute the hyper-parameter values, we use a randomized search method [19] with 100 iterations that samples the hyper-parameters according to an exponential distribution (with scale 100). To simplify our retrieval model, we consider all articles to have the same set of weights for their fields.

The queries describe patient cases using the fields mentioned in Table 6.6. Table 6.1 represents an example of queries used in this task. The gene field contains gene names, such as "BRAF[15]" and "NRAS[16]". In 10 patient cases, the AA mutation codes of the genes were also available, e.g., "V600E[17]" and "Q61K[18]". In 13 patient cases, instead of AA mutation codes, the types of mutation (described by "amplification", "deletion", "fusion", "fusion transcript", "inactivating", and "loss" in the query) were provided in the queries. In 8 patient cases, only the names of mutated genes were provided. Of 30 patient cases, which were described in the form

[15] B-Raf proto-oncogene, serine/threonine kinase
[16] NRAS proto-oncogene, GTPase
[17] BRAF c.1799T>A
[18] NRAS c.181C>A

of queries, 50% were male, and 50% were female. These cases were created by precision oncologists at the MD Anderson Cancer Center and the OHSU Knight Cancer Institute [90]. We present a histogram of diseases (cancer types) for the patients described in the query set of our training data in Figure 6.6(a). From this figure, we can see that patients with 17 types of cancers were included. The patients with lung related cancer types were in the majority, representing 7 of 30 patients. On the other hand, the age distribution of patients with respect to their gender, demonstrated in Figure 6.6(b), demonstrates that all patients described in the query set ranged

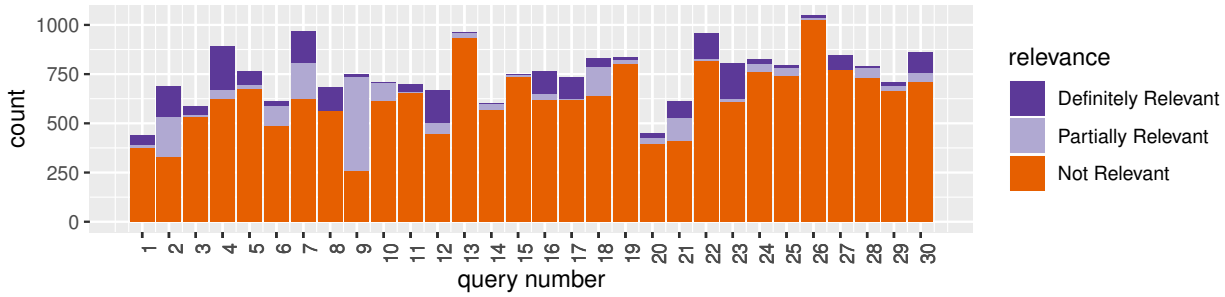


Figure 6.5: Counts of relevance judgments of the queries in the training data for the three levels of relevance: “Definitely Relevant”, “Partially Relevant” and “Not Relevant”.

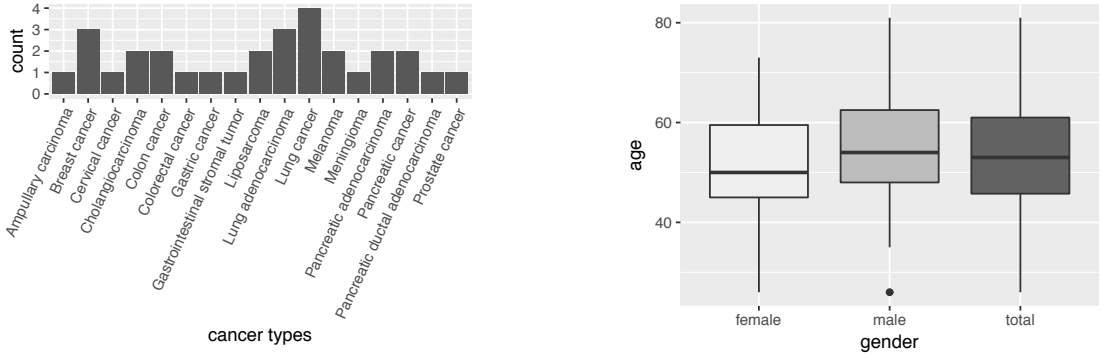
Table 6.6: Properties of fields in the 30 patient cases described in the form of queries.

Fields	Properties
Disease	17 cancer types, such as “cholangiocarcinoma”
Gene	Up to 3 genes, such as “KRAS”
Mutation	In 10 queries, specific mutation codes, such as “Q61K”, are specified
Age	Average: 52.6 years, standard deviation: 13.5, and range: 26-81 years
Gender	50% female and 50% male
Other	The other factors include secondary diseases (such as “Type II Diabetes”), symptoms (such as “hypertension”), and surgical procedures (such as “Whipple”)

in age between 26 and 81 years, with a mean age of 52.6 years (mean age of female patients, 51.2 years; mean age of male patients, 54 years) and a standard deviation of 13.54.

Baselines and Variations of the Model

We adopted the best-performing methods in TREC-PM 2017 [37] and TREC-CDS 2015 [13] as two of the baselines. We considered INTGR (INTeGrating semantic and statistical concepts for medical query expansion) [12] as another baseline because it is an optimized method to integrate knowledge bases to perform query expansion.



(a) (b)

Figure 6.6: (a) histogram of cancer types and (b) distributions of ages with respect to the gender of patients in the query set used for tuning the parameters of BPM.

To study the effects of the three knowledge bases mentioned in Section 6.2 in generating the list of candidate expansion concepts for query expansion, we considered the following four variations of BPM:

- **UMLS-BPM**: In this variation, the list of candidate expansion concepts contains only concepts from the table of relationships in UMLS (i.e., MRREL.RRF). In other words, we only considered concepts that are related to the UMLS concepts in the original query according to this table.
- **DGIdb-BPM**: In this variation, we added a concept to the list of candidate expansion concepts if it had an interaction with the concepts in the query according to DGIdb table of interactions of drugs and genes.

- **RM-BPM**: In this variation, we added concepts to the list of candidate expansion concepts if they were highly ranked by the relevance feedback (RM) model [52]. We ran the original query after concatenating all of its fields into a single free text query, obtained the top 40 documents, and considered the top 40 highly ranked unigram and bigram concepts extracted from these documents as candidate concepts.
- **Wiki-BPM**: This baseline is similar to RM-BPM, but instead of running original queries on a PubMed collection to obtain the top-ranked documents, we ran the queries on a Wikipedia dump (01/01/2018) and ranked candidate expansion concepts by using the RM model. Our main goal in considering Wiki-BPM is to evaluate the effectiveness of top-ranked documents from PubMed in comparison to those retrieved from a general-purpose collection of documents.

Baselines and Variations of the Model

Setup. We use the following four evaluation metrics in comparing BPM with the baselines and its four variations in Table 6.7.

- infNDCG (inferred normalized discounted cumulative gain) [121]: infNDCG uses sampling techniques to estimate NDCG (normalized discounted cumulative gain) [65] by incorporating graded relevance judgments with missing values. NDCG is derived by normalizing the Discounted Cumulative Gain (DCG) measure, and DCG (discounted cumulative gain) is obtained from total accumulated relevancy gains discounted by giving higher weights to the documents with higher ranks.
- P@10 (precision at 10) [65]: P@10 is the percentage of relevant documents in top 10 retrieved documents.

- R-prec (R-precision) [65]: Given R as the number of relevant documents for the query, R-prec is defined as precision at R.

Result. Table 6.7 shows that BPM has a statistically significant improvement over the best-performing baseline (UTDHLTFF [37]) when BPM uses all three knowledge bases described in Section 6.2 to generate the list of candidate expansion concepts. Because UTDHLTFF [37] uses a similar set of knowledge bases, this improvement is an indication of the effectiveness of our Bayesian approach in ranking the candidate concepts for query expansion. This table shows that without using all three knowledge bases as a source of generating a list of candidate expansion concepts, the improvement in the performance of BPM over UTDHLTFF [37] is not significant.

Effect of Knowledge Bases on Generating the List of Candidate Expansion Concepts

Table 6.7 shows that the concepts from top-ranked documents improve the quality of the query more than the concepts obtained from the UMLS table of relationships and the DGIdb table of interactions of drugs and genes. This difference is potentially due to (1) the noise in the UMLS relationship table and DGIdb and (2) the fact that the top-ranked documents are more dependent

Table 6.7: Comparison of BPM with state-of-the-art baselines using the TREC-PM 2017 query set. The statistical significance of BPM in comparison to UTDHLTFF [37] according to a one-sided Fisher's randomization test computed at the 95% significance level is shown by * in this table.

Methods	infNDCG	R-prec	P@10
WSU-IR [13]	0.3853	0.2682	0.5937
INTGR [12]	0.4021	0.2739	0.6010
UTDHLTFF [37]	0.4593	0.2987	0.6172
UMLS-BPM	0.4507	0.2952	0.6166
DGIdb-BPM	0.4556	0.2970	0.6191
RM-BPM	0.4624	0.2937	0.6135
Wiki-BPM	0.4611	0.2996	0.6188
BPM	**0.4837***	**0.3160***	**0.6292***

on the given patient case. In this section, we provide a deeper analysis of the effect of each knowledge base on the performance of BPM.

Setup. In Table 6.7, we present the performance of BPM by varying the weights of expansion concepts from different resources (i.e., ω_x in (6.16)) in computing the relevance score of a collection document (D) given an expanded query ($\tilde{Q}$). In this figure, the weights ω_{UMLS}, ω_{RM}, and $\omega_{DGIdb} = 1 - \omega_{UMLS} - \omega_{RM}$ are the weights of concepts obtained from UMLS, top-ranked documents from PubMed, and DGIdb, respectively. If a concept is obtained from multiple concept resources, we remove it from the list of expansion concepts to avoid ambiguity in our analysis.

Result. From Table 6.7, we observe that the concepts that BPM extracted from top-ranked documents provide a significantly better improvement than the concepts from other resources. More specifically, this table reveals that BPM performs best when the concepts from top-ranked documents have the highest weight, which highlights the importance of adjusting the list of candidate concepts by employing resources (such as top-ranked documents) that are more dependent on the query itself.

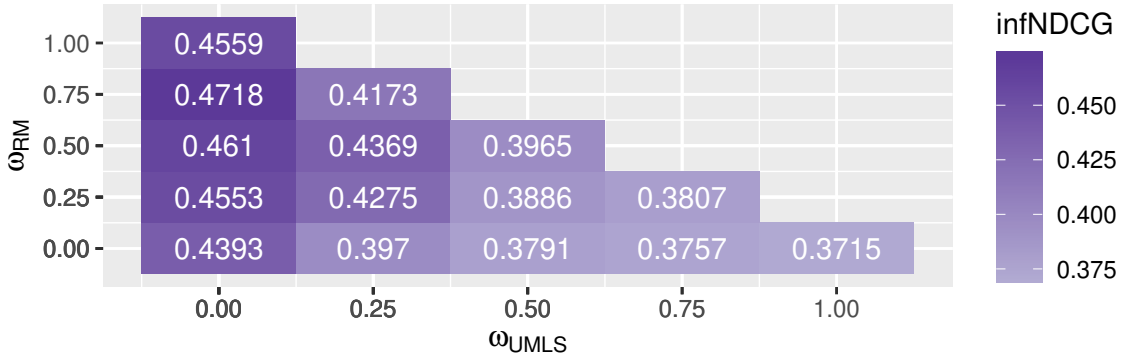


Figure 6.7: Performance of BPM in terms of infNDCG in the case of generating the list of candidate expansion concepts by utilizing the concept resources UMLS, DGIdb and top-ranked documents (RM) from PubMed with different set of weights for their corresponding expansion concepts. The

weight of DGIdb (ω_{DGIdb}) is obtained from the weight of UMLS (ω_{UMLS}) and RM (ω_{RM}) as $\omega_{DGIdb} = 1 - \omega_{UMLS} - \omega_{RM}$. These weights are shown by ω_x in (6.16).

Query Level Analysis

Setup. We analyze the performance of BPM in comparison with that of UTDHLTFF (the best-performing method in TREC-PM 2017) in Figure 6.8 by comparing the improvement of BPM over its baseline at the query level. The thirty queries shown in this figure are from the query set employed in TREC-PM 2017[19].

Result. Figure 6.8 demonstrates that in 63.33% of the queries, BPM outperforms its baseline (UTDHLTFF [37]). By studying the best-performing queries, we understand that BPM has the advantage of finding documents that do not have any of the terms explicitly mentioned in the query. This shows that, by utilizing a Bayesian approach to expand the queries with their related concepts, BPM can fill the vocabulary gap between queries and their relevant documents in the collection. If we define difficult query as a query that has an infNDCG value lower than 0.1 given UTDHLTFF as the baseline retrieval method, BPM often has lower performance on difficult queries than its baseline because it relies on top-ranked documents, which often do not provide reliable expansion concepts when the query is difficult [120].

Effect of Query Features

Setup. To understand the effect of each query feature, we examine the performance of BPM by varying the weight of features $\mathbf{F}_q$. We present the results of this experiment in Figure 6.9.

[19] http://www.trec-cds.org/topics2017.xml

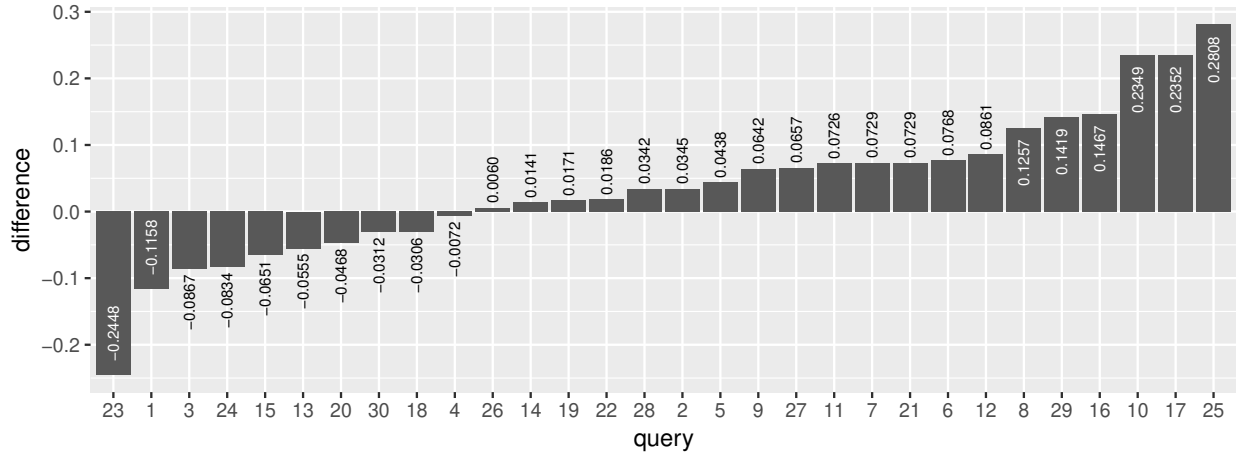
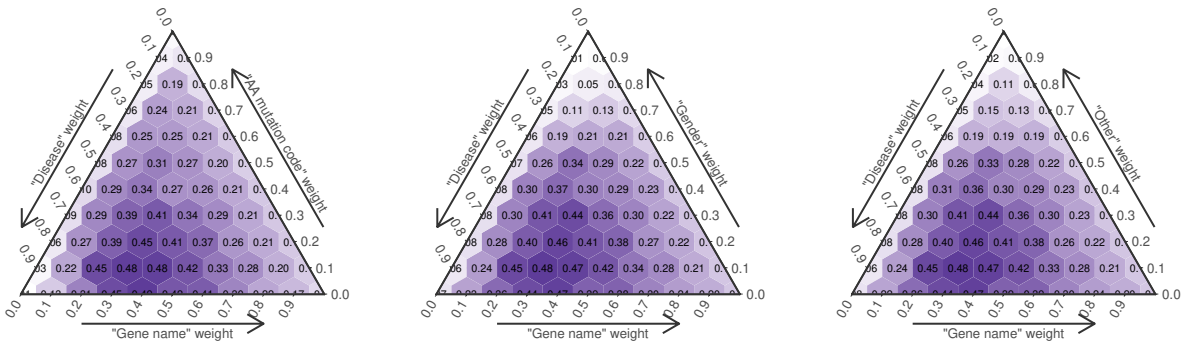


Figure 6.8: Query-level analysis of BPM over the best-performing baseline (UTDHLTFF). This figure shows the performance improvement of BPM over UTDHLTFF in terms of infNDCG on the query level.

Result. Figure 6.9 shows that the query features “Disease” and “Gene name” have higher importance than all other features in the query. The feature “Gender” has the least importance in comparison to the other features in the query. Although the query feature “AA mutation code” provides information regarding the patient cases, it has less importance than “Gene name” and “Disease” in our retrieval model. In Figure 6.10, we show the percentage of documents in the training data that contain the queries' AA mutation codes and are either relevant or nonrelevant. We observe from Figure 6.10 that “AA mutation code” in the queries tends to occur more in nonrelevant documents than in relevant ones and therefore can cause the topic-drift problem in our query expansion model. We can conclude from Figure 6.9 that by using only the features “Disease” and “Gene name”, we can simplify our retrieval model with a negligible decrease in retrieval performance.



(a) (b) (c)

Figure 6.9: Weights of the most important query features (i.e., "Disease" and "Gene name") in comparison to the query features "AA mutation code" (shown in Figure 6.9(a)), "Gender" (shown in Figure 6.9(b)) and "Other" (shown in Figure 6.9(c)). The weight of the l-th query feature is shown by ω_l in (6.15).

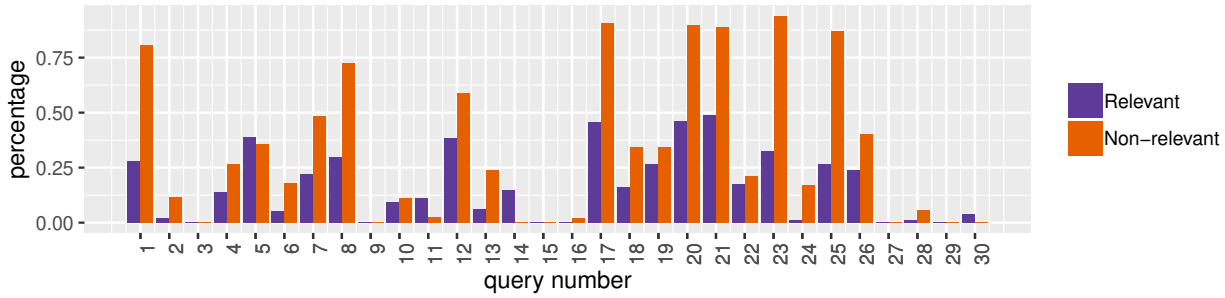


Figure 6.10: Percentage of documents in the training data that contain queries' AA mutation codes and are either relevant or nonrelevant. This figure demonstrates that the queries' AA mutation codes tend to occur more in nonrelevant documents than in relevant ones. Therefore, using this query field in a retrieval model can cause a decrease in the accuracy of retrieved documents. In addition, using this field in a query expansion model can cause the topic-drift problem.

Table 6.8: Performance of BPM with respect to different values of unigram weights (λ_T) in SDM.

	infNDCG	R-prec	P@10
BPM ($\lambda_T = 0$)	0.1783	0.0914	0.2132
BPM ($\lambda_T = 0.25$)	0.2416	0.1332	0.2865
BPM ($\lambda_T = 0.5$)	0.3467	0.2755	0.4653
BPM ($\lambda_T = 0.75$)	**0.4902**	**0.3318**	**0.6410**
BPM ($\lambda_T = 1$)	0.4473	0.2841	0.6011

Effect of Using the Sequential Dependence Model to Represent Concepts

Setup. We examine the weight of unigrams (λ_T) in comparison to the weights of ordered and unordered bigrams (λ_O and λ_U) in the SDM described in sections 6.2 and 6.2 and formulated in (6.17). Table 6.8 presents the performances of BPM for different values of weights of unigrams in the SDM when the weights of ordered and unordered bigrams are assumed to be the same (i.e., $\lambda_U = \lambda_O = 1 - \lambda_T$).

Result. Table 6.8 demonstrates that the unigrams have the most critical role in SDM of medical concepts. In other words, we observe that when $\lambda_T = 0.750$, $\lambda_U = 0.125$, and $\lambda_O = 0.125$, the performance of BPM is better than that achieved when lower or higher values of λ_T are used.

Success and Failure Analysis

Setup. We examine the performance of BPM by analyzing its successes and failures in comparison to the best-performing baseline UTDHLTFF [37]. Table 6.9 represents the best- and worst-performing queries.

Result. In the best- and worst-performing queries, the disease names are lung adenocarcinoma and breast cancer and gene names are MET[20] and PTEN[21]. We observe that the best- and worst-performing queries are those that provide the most and fewest useful concepts in their list of expansion concepts, respectively. More specifically, we observe that the list of expansion concepts obtained from the RM has significantly higher quality in the case of the best-performing query than in the case of the worst-performing query. In the best performing query,

[20] MET proto-oncogene, receptor tyrosine kinase
[21] phosphatase and tensin homolog

the improvement in the quality of query is mainly due to the addition of the concepts "EGFR[22]" and "NSCLC[23]" which often appear in the relevant documents. BPM extracts both of these two concepts from top-ranked documents. In the worst-performing query, the main reason of decline in the retrieval performance is due to the addition of the expansion concept "heart failure" to the query. This concept appears in a number of top-ranked documents such as PMID: 10321507 (https://www.ncbi.nlm.nih.gov/pubmed/10321507) but it does not appear in the articles judged as relevant. These results reveal that depending on the query, the quality of the concepts extracted from top-ranked documents can be significantly different. Therefore, having resources other than top-ranked documents can increase the effectiveness of query expansion method on average for all the queries.

Table 6.9: The best- and worst-performing queries for BPM in comparison to the best performing baseline UTDHLTFF [37].

number	25
disease	Lung Adenocarcinoma
gene	MET Amplification
demographic	48-year-old Male
other	Emphysema

(a) Best-performing Query

number	23
disease	Breast Cancer
gene	PTEN Loss
demographic	54-year-old Female
other	Congestive Heart Failure

(b) Congestive Heart Failure

Qualitative Analysis

Setup. We provide a qualitative analysis of BPM given the example query in Table 6.1. "PIK3CA (E545K)" is the gene mutation, and "gastric cancer" is the type of cancer described in this example query. The patient is a 54-year-old male who suffers from depression according to the query. In our Bayesian approach, we perform three steps: generating the list of candidate

[22] epidermal growth factor receptor
[23] non-small-cell lung carcinoma

concepts for query expansion, computing prior probabilities, and computing posterior probabilities. In this qualitative analysis, for the sake of simplicity, BPM generates the list of candidate expansion concepts by selecting only unigram concepts that exist in top-ranked documents for the original query (obtained using the RM model). In this section, we also consider the set of features Fs containing features of only types "Primary site" or "Primary histology" extracted from the COSMIC knowledge base.

Result.

Step I - generating the list of candidate expansion concepts: BPM uses the concepts shown in Figure 6.11, which were obtained by using the RM model from top-ranked documents, as the list of candidate expansion concepts. To examine the effectiveness of each candidate expansion concept on the retrieval performance of BPM, we expand the query with each individual concept and measure the amount of improvement this query expansion provides. We categorize these concepts based on their individual effects on retrieval performance as follows:

1. concepts with positive effects, such as "gastroesophageal", "amplification", and "stomach";
2. concepts with no effect, such as "patient", "study", and "cell"; and
3. concepts with negative effects, such as "carcinoma", "tumor", and "AKT[24]".

[24] Protein kinase B

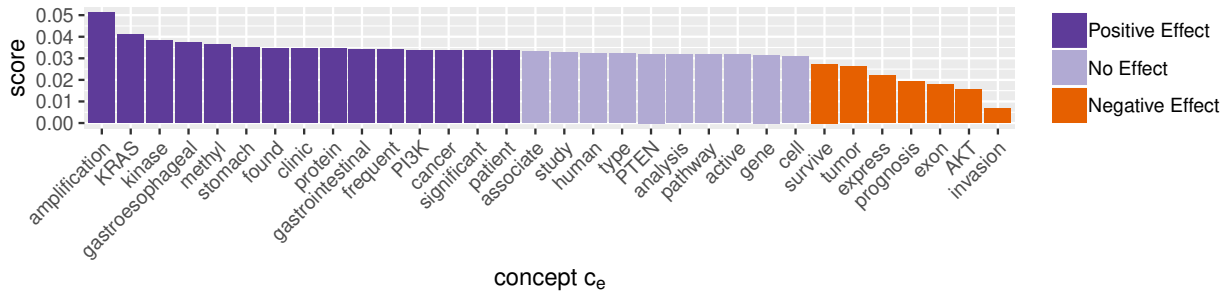


Figure 6.11: An illustration of the list of candidate concepts and their effectiveness scores for the example query shown in Table 6.1. We obtain these concepts by using the RM model from top-ranked documents. The effectiveness scores of the candidate expansion concepts are computed by expanding the query with only one of these concepts and computing the improvement in the value of infNDCG of the retrieved documents.

We observe from this figure that the concepts with no effect are often general concepts that can occur in any medical article, while concepts that have positive or negative effects are those that tend to be more discriminative (i.e., occur in more limited medical articles). Figure 6.11 shows that approximately half of the concepts in this list have no or a negative effect on the quality of the expanded query. This figure also reveals that the concepts with positive effects tend to have higher relevance to the given query.

Step II - **computing the prior probabilities**: To compute the prior probabilities $p(c_e|g)$ from (6.7) BPM first computes the probability of feature f extracted from COSMIC being related to gene g, i.e., $p(f|g)$, and the probability of candidate expansion concept c_e being related to feature f and gene g, i.e., $p(c_e|g,f)$.

BPM computes $p(f|g)$ from (6.8) by counting the number of cancerous patients with mutated gene g (i.e., N_g) and the number of cancerous patients with mutated gene g that have feature f (i.e., $N_{f,g}$) in COSMIC. An illustration of computed $p(f|g)$ for the example query in Table 6.1 is shown in Figure 6.12 for the case of having the feature types “Primary site” and “Primary histology”. For simplicity, we only consider these two feature types in the experiments in this

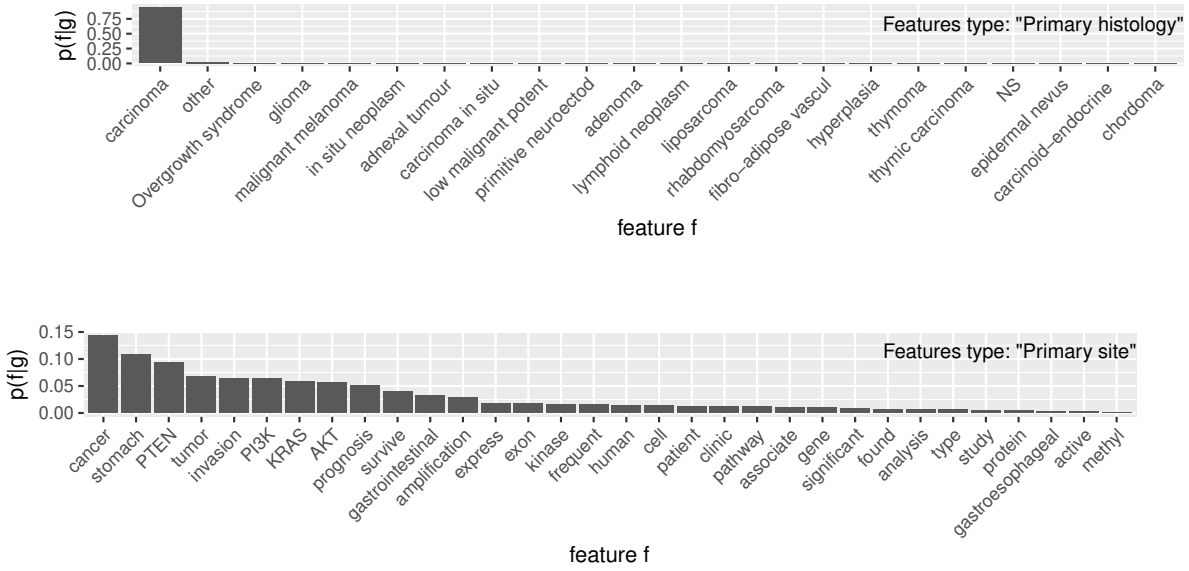


Figure 6.12: An illustration of prior probability $p(f|g)$ for feature f ("Primary site" and "Primary histology") being related to the mutated gene g ="PIK3CA". To compute this probability, $N_{f,g}$ is normalized by $N_g = 2737$, which is the number of patient cases that have the mutated gene "PIK3CA" in the COSMIC knowledge base.

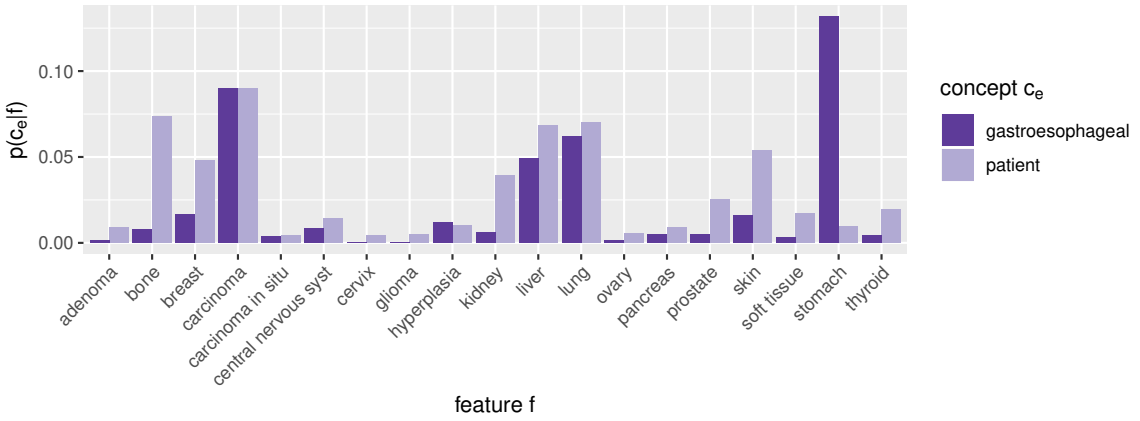


Figure 6.13: An illustration of the probability of candidate expansion concept c_e being related to feature f (i.e., $p(c_e|f)$) for the mutated gene "PIK3CA" and features of type "Primary site" and "Primary histology". Not all the values of features are shown in this figure for the sake of visibility. In this example, only two concepts, "gastroesophageal" and "patient", are studied. The probabilities in this figure are obtained by normalizing $|R_{c_e,f}|$ by the number of documents in the collection that contain concept c_e (i.e., $|R_{c_e}|$). $|R_{c_e}|$ equals 19600 and 5421011 for the concepts "gastroesophageal" and "patient", respectively.

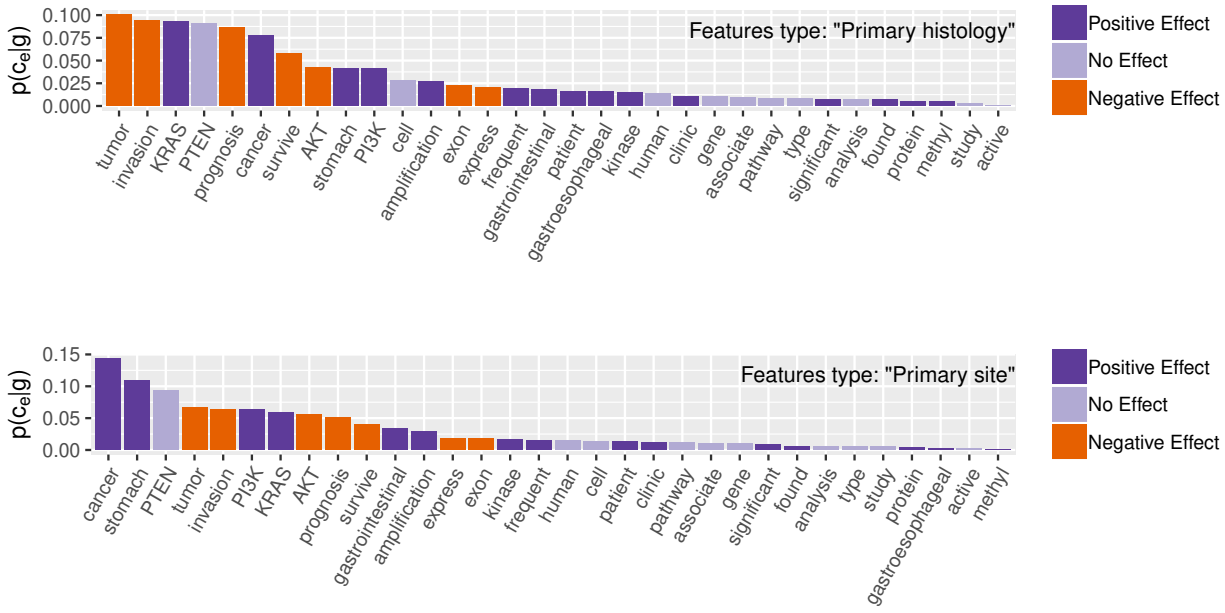
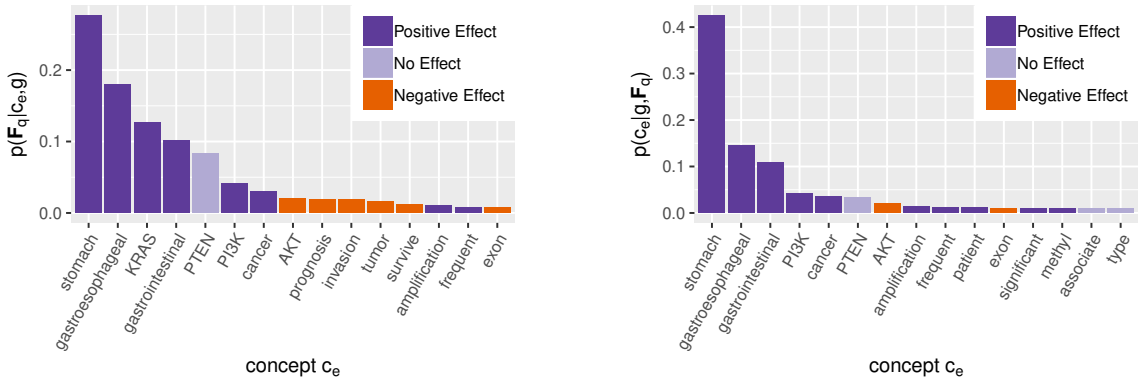


Figure 6.14: An illustration of the estimated values for the prior probability $p(c_e|g)$ for the mutated gene g ="PIK3CA" and candidate expansion concepts c_e in the case of using the features of type "Primary site" and "Primary histology".



(a) (b)

Figure 6.15: An illustration of the estimated values for the (a) likelihood $p(\mathbf{F}_q|c_e, g)$ and (b) posterior probability $p(c_e|g, \mathbf{F}_q)$ for the mutated gene g ="PIK3CA" and candidate expansion concepts c_e extracted from top-ranked documents. Only the top 15 concepts are shown in this figure.

section. This figure implies that according to COSMIC, in the majority of the cancerous patient cases, the primary sites affected by the gene mutation "PIK3CA (E545K)" are breast (in 33.54% of

the cases) and large intestine (in 26.85% of the cases). Although the patient case described in the example query has gastric cancer with the mutated gene "PIK3CA", according to COSMIC, the stomach is the primary site affected by this mutated gene in only 2.81% of cancerous patient cases. On the other hand, the primary histology for the mutated gene "PIK3CA" is "carcinoma" in 94.06% of the patient cases.

As the next step in computing the prior probability, BPM computes the probability of concept c_e being related to mutated gene g and feature f, i.e., $p(c_e|g,f)$, from (6.9) by using the medical articles in PubMed as the knowledge base. To do so, BPM computes the cooccurrence of ngram representations of c_e, g, and f in the collection of medical articles in PubMed. An illustration of computed values of $p(c_e|g,f)$ for the concepts "gastroesophageal" and "patient" is shown in Figure 6.13.

By using (6.7), BPM computes the prior probabilities $p(c_e|g)$ from the probabilities $p(f|g)$ and $p(c_e|g,f)$. Figure 6.14 highlights that by using the computed prior probabilities, BPM is able to distinguish the discriminative concepts (often with positive or negative effects) from general concepts (often with no effect).

Step III - computing the likelihoods and posterior probabilities given evidence from the query: At this step, BPM incorporates information regarding a patient case from the given query in computing the posterior probabilities. To do so, as illustrated in Figure 6.15(a), BPM first computes the likelihood $p(\mathbf{F}_q|c_e,g)$. This figure suggests that BPM gives the highest scores at this stage to the concepts that are more related to the patient cases described in the query. By using the likelihood and prior probability, BPM computes the posterior probabilities illustrated in Figure 6.15(b). By choosing concepts that have posterior probabilities $p(c_e|g,\mathbf{F}_q)$ above a

threshold, BPM selects candidate concepts as the expansion concepts. We can observe from Figure 6.15(b) that the majority of selected concepts have positive effects on the retrieval performance of BPM.

CHAPTER 7 CONCLUSIONS

In this dissertation, we presented a concept representation method and an optimization technique to jointly determine the weights of statistical and semantic concepts from different sources. Our proposed methods represent CDS queries using statistical and semantic concepts from the query, top retrieved documents and knowledge bases. Our work logically extends previous research, which focused only on studying the utility of statistical query concepts [16], semantic query concepts [15], statistical and semantic query concepts [27], statistical [70, 17] and semantic [101] concepts from the query and top retrieved documents for query expansion. Experiments using a collection of PubMed articles and TREC Clinical Decision Support (CDS) track queries indicate that the proposed method significantly outperforms state-of-the-art baselines for ad hoc and medical IR.

We also presented a two-stage method for sequential selection of effective concepts for query expansion from the concept graph. We formulated an optimization problem with the objective of evaluating the least possible number of candidate concepts needed to ensure a given precision of retrieval results. In the first stage of the proposed method, the candidate concepts are sorted using a number of computationally inexpensive features. This sorting is utilized in the second stage to sequentially select expansion concepts by using computationally expensive features. Experimental evaluation using TREC collections indicates that the proposed method outperforms state-of-the-art baselines, which instead of minimizing the number of evaluated concepts, aim to minimize the number of selected concepts or maximize a concept quality measure. We also found out that our method and the baselines produce more accurate results using ConceptNet-based than the collection-based concept graph HAL. We believe that applying

our method to the case of entity-based queries and knowledge graphs is an interesting future direction for extending this work.

Finally, we proposed an information retrieval method for a clinical decision support system in the precision medicine paradigm. Through a Bayesian approach, our method incorporates information gathered from multiple knowledge bases including a collection of biomedical articles in PubMed and Catalog of Somatic Mutations in Cancer (COSMIC). Our Bayesian approach for query expansion improves the retrieval accuracy by discovering related concepts that can fill the vocabulary gap between a medical query and its relevant documents in the collection. Since in a precision medicine paradigm, the mutated gene mentioned in a query provides critical information regarding a patient case, our method first utilizes knowledge bases to rank a list of candidate concepts for query expansion based on their relatedness to the mutated gene in the query. Next, our method utilizes the other information mentioned in the query to update its prior belief regarding the relatedness of a candidate expansion concept to the query. We performed experiments on the 2017 TREC-PM dataset and observed that our method significantly outperforms state-of-the-art baselines.

REFERENCES

[1] Health expenditure @ONLINE, 2016.

[2] I. Alonso and D. Contreras. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An umls approach. Expert Systems with Applications, 44:386–399, 2016.

[3] R. Anand and A. Kotov. Improving difficult queries by leveraging clusters in term graph. In Proceedings of the 11th AIRS, pages 426–432, 2015.

[4] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Using literature and data to learn bayesian networks as clinical models of ovarian tumors. Artificial Intelligence in medicine, 30(3):257–281, 2004.

[5] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In Proceedings of the AMIA Symposium, page 17. American Medical Informatics Association, 2001.

[6] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229–236, 2010.

[7] A. R. Aronson and T. C. Rindflesch. Query expansion using the umls metathesaurus. In Proceedings of the AMIA Annual Fall Symposium, page 485. American Medical Informatics Association, 1997.

[8] S. J. Aronson and H. L. Rehm. Building the foundation for genomics in precision medicine. Nature, 526(7573):336, 2015.

[9] A. M. Association. Current procedural terminology: CPT. American Medical Association, 2007.

[10] A. Bair, L. Brown, L. Pugh, L. Borucki, and D. Spatz. Taking a bite out of crisp. strategies on using and conducting searches in the computer retrieval of information on scientific projects database. Computers in nursing, 14(4):218–24, 1995.

[11] S. Balaneshin-kordan and A. Kotov. An empirical comparison of term association and knowledge graphs for query expansion. In Proceedings of the 38th ECIR, pages 761–767, 2016.

[12] S. Balaneshin-kordan and A. Kotov. Optimization method for weighting explicit and latent concepts in clinical decision support queries. In Proceedings of the 2nd ACM ICTIR, 2016.

[13] S. Balaneshin-kordan, A. Kotov, and R. Xisto. WSU-IR at TREC 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In Proceedings of TREC'15, 2015.

[14] K. A. Balaneshinkordan, Saeid and R. Xisto. Wsu-ir at trec 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In Proc. 24th Text Retrieval Conference (TREC 2015). National Institute of Standards and Technology (NIST), 2015.

[15] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 491–498. ACM, 2008.

[16] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In Proceedings of the third ACM international conference on Web search and data mining, pages 31–40. ACM, 2010.

[17] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 605–614. ACM, 2011.

[18] B. G. Bentsen. International classification of primary care. Scandinavian journal of primary health care, 4(1):43–50, 1986.

[19] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb):281–305, 2012.

[20] A. Blake and A. Zisserman. Visual reconstruction, volume 2. MIT press Cambridge, 1987.

[21] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl 1):D267–D270, 2004.

[22] R. C. Bodner and F. Song. Knowledge-based approaches to query expansion in information retrieval. Springer, 1996.

[23] E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (meddra). Drug Safety, 20(2):109–117, 1999.

[24] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discourse. Discourse Processes, 25(2-3):211–257, 1998.

[25] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In Proceedings of the 31st ACM SIGIR, pages 243–250, 2008.

[26] S. Choi and J. Choi. Snumedinfo at trec cds track 2014: medical case-based retrieval task. Technical report, DTIC Document, 2014.

[27] S. Choi, J. Choi, S. Yoo, H. Kim, and Y. Lee. Semantic concept-enriched dependence model for medical information retrieval. Journal of biomedical informatics, 47:18–27, 2014.

[28] F. S. Collins and H. Varmus. A new initiative on precision medicine. New England Journal of Medicine, 372(9):793–795, 2015.

[29] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. Machine learning, 9(4):309–347, 1992.

[30] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pages 365–374. ACM, 2014.

[31] F. Diaz. Pseudo-query reformulation. In Proceedings of the 38th ECIR, pages 521–532, 2016.

[32] D. Eichmann, M. E. Ruiz, and P. Srinivasan. Cross-language information retrieval with the umls metathesaurus. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 72–80. ACM, 1998.

[33] E. Erdem, E. Kocabas, H. Taylan Sekeroglu, O¨. Ozgur, M. Yagmur, and T. R. Ersoz. Crystalline-like keratopathy after intravenous immunoglobulin therapy with incomplete kawasaki disease: Case report and literature review. Case Reports in Ophthalmological Medicine, 2013, 2013.

[34] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic acids research, pages D945–D950, 2010.

[35] J. I. Garcia-Gathright, F. Meng, and W. Hsu. Ucla at trec 2014 clinical decision support track: Exploring language models, query expansion, and boosting. Technical report, DTIC Document, 2014.

[36] J. Gobeill, A. Gaudinat, E. Pasche, and P. Ruch. Full-texts representations with medical subject headings, and co-citations network reranking strategies for trec 2014 clinical decision support track. Technical report, DTIC Document, 2014.

[37] T. Goodwin, M. Skinner, and S. Harabagiu. Utd hltri at trec 2017: Precision medicine track. In Proceedings of TREC-PM, pages 1–9, 2017.

[38] M. Griffith, O. L. Griffith, A. C. Coffman, J. V. Weible, J. F. McMichael, N. C. Spies, J. Koval, I. Das, M. B. Callaway, J. M. Eldred, et al. Dgidb: mining the druggable genome. Nature methods, page 1209, 2013.

[39] F. Hasibi, K. Balog, and S. E. Bratsberg. Entity linking in queries: tasks and evaluation. In Proceedings of the 2015 ICTIR, pages 171–180, 2015.

[40] W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In Proceedings of the AMIA Symposium, page 344. American Medical Informatics Association, 2000.

[41] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett. The unified medical language system. Journal of the American Medical Informatics Association, 5(1):1–11, 1998.

[42] A. Jimeno-Yepes and A. R. Aronson. Improving an automatically extracted corpus for umls metathesaurus word sense disambiguation. Procesamiento del lenguaje natural, 45:239–242, 2010.

[43] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, et al. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research, pages D1071–D1078, 2014.

[44] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. The Australasian medical journal, 5(9):482, 2012.

[45] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. Information retrieval as semantic inference: A graph inference model applied to medical search. Information Retrieval Journal, pages 1–32, 2016.

[46] B. Koopman, G. Zuccon, A. Nguyen, D. Vickers, L. Butt, and P. Bruza. Exploiting snomed ct concepts & relationships for clinical information retrieval: Australian e-health research centre and queensland university of technology at the trec 2012 medical track. Technical report, DTIC Document, 2012.

[47] A. Kotov and C. Zhai. Interactive sense feedback for difficult queries. In Proceedings of the 20th ACM international conference on Information and knowledge management, pages 163–172. ACM, 2011.

[48] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In Proceedings of the fifth

ACM international conference on Web search and data mining, pages 403–412. ACM, 2012.

[49] H. Lang, D. Metzler, B. Wang, and J.-T. Li. Improved latent concept expansion using hierarchical markov random fields. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 249–258. ACM, 2010.

[50] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. Machine learning, 81(1):53–67, 2010.

[51] V. Lavrenko and W. B. Croft. Relevance based language models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 120–127. ACM, 2001.

[52] V. Lavrenko and W. B. Croft. Relevance-based language models. In ACM SIGIR Forum, pages 260–267, 2017.

[53] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning, pages 4–15. Springer, 1998.

[54] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 3–12. Springer-Verlag New York, Inc., 1994.

[55] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring conceptual relationships to improve medical records search. In Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pages 1–8, 2013.

[56] J. Lin and D. Demner-Fushman. The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine. In Proceedings of SIGIR'06, pages 99–106, 2006.

[57] H. Liu and P. Singh. Conceptnet – a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226, 2004.

[58] Z. Liu and W. W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Information Retrieval, 10(2):173–202, 2007.

[59] P. Lopez-Garcia, M. Oleynik, Z. Kasac, and S. Schulz. Trec 2017 precision medicine - medical university of graz. In Proceedings of TREC-PM, pages 1–12, 2017.

[60] H. J. Lowe and G. O. Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. Jama, 271(14):1103–1108, 1994.

[61] Z. Lu, W. Kim, and W. J. Wilbur. Evaluation of query expansion using mesh in pubmed. Information retrieval, 12(1):69–80, 2009.

[62] C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In Proceedings of the ACM CIKM, pages 341–350, 2007.

[63] A. A. Mahmood, G. Li, S. Rao, P. McGarvey, C. Wu, S. Madhavan, and K. Vijay- Shanker. Ud gu biotm at trec 2017: Precision medicine track. In Proceedings of TREC-PM, pages 1–9, 2017.

[64] C. D. Manning and H. Schu¨tze. Foundations of statistical natural language processing. MIT press, 1999.

[65] P. R. Manning, Christopher D. and H. Schutze. Introduction to information retrieval, volume 1. Cambridge University Press, 2008.

[66] A. T. McCray. The umls semantic network. In Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care, pages 503–507. American Medical Informatics Association, 1989.

[67] A. T. McCray and S. J. Nelson. The representation of meaning in the umls. Methods of information in medicine, 34(1-2):193–201, 1995.

[68] C. McDonald, S. Huff, J. Suico, and K. Mercer. Logical observation identifiers names and codes (loinc) users' guide. Indianapolis: Regenstrief Institute, 2004.

[69] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 472–479. ACM, 2005.

[70] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 311–318. ACM, 2007.

[71] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. Information Retrieval, 10(3):257–274, 2007.

[72] D. A. Metzler, W. B. Croft, and A. McCallum. Direct maximization of rank-based metrics for information retrieval. Technical report, Center for Intelligent Information Retrieval, 2005.

[73] A. Miller. Subset selection in regression. CRC Press, 2002.

[74] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 206–214. ACM, 1998.

[75] H. Mobahi and J. W. Fisher III. On the link between gaussian homotopy continuation and convex envelopes. In Energy Minimization Methods in Computer Vision and Pattern Recognition, pages 43–56. Springer, 2015.

[76] W. Morgan, W. Greiff, and J. Henderson. Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In Proceedings of NAACL-HLT'04, pages 93–96, 2004.

[77] A. Mourao, F. Martins, and J. Magalhaes. Novasearch at trec 2014 clinical decision support track. Technical report, DTIC Document, 2014.

[78] X. Mu and K. Lu. Towards effective genomic information retrieval: The impact of query complexity and expansion strategies. Journal of Information Science, 36(2):194–208, 2010.

[79] X. Mu and K. Lu. Improving umls metathesaurus query expansion based on the query specificity and length. 2012.

[80] F. B. Nardon and L. A. Moura. Knowledge sharing and information integration in healthcare using ontologies and deductive databases. Medinfo, 11(Pt 1):62–6, 2004.

[81] R. Navigli. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41(2):10, 2009.

[82] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore. Normalized names for clinical drugs: Rxnorm at 6 years. Journal of the American Medical Informatics Association, 18(4):441–448, 2011.

[83] V. Nguyen, S. Karimi, S. Falamaki, and C. Paris. Benchmarking clinical decision support search. arXiv preprint arXiv:1801.09322, 2018.

[84] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In Proceedings of the 29th ACM SIGIR, pages 91–98, 2006.

[85] P. Norvig. Marker passing as a weak method for text inferencing. Cognitive Science, 13(4):569–620, 1989.

[86] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In Proceedings of the 21st ACM SIGIR, pages 275–281, 1998.

[87] D. Rao, P. McNamee, and M. Dredze. Entity linking: Finding extracted entities in a knowledge base. In Multi-source, Multilingual Information Extraction and Summarization, pages 93–115. Springer, 2013.

[88] T. C. Rindflesch and A. R. Aronson. Ambiguity resolution while mapping free text to the umls metathesaurus. In Proceedings of the Annual Symposium on Computer Application in Medical Care, page 240. American Medical Informatics Association, 1994.

[89] K. Roberts, D. Demner-Fushman, E. M. Voorhees, and W. R. Hersh. Overview of the trec 2016 clinical decision support track. In Proceedings of Text Retrieval Conference (TREC), 2016.

[90] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, and S. Pant. Overview of the trec 2017 precision medicine track. In Proceedings of TREC-PM, pages 1–13, 2017.

[91] K. Roberts, M. S. Simpson, E. Voorhees, and W. R. Hersh. Overview of the trec 2015 clinical decision support track. Proceedings of TREC'15, 2015.

[92] R. W. Schafer. What is a Savitzky-Golay filter? [lecture notes]. IEEE Signal Processing Magazine, 28(4):111–117, 2011.

[93] W. Shen and J.-Y. Nie. Is concept mapping useful for biomedical information retrieval? In Experimental IR Meets Multilinguality, Multimodality, and Interaction, pages 281–286. Springer, 2015.

[94] W. Shen, J.-Y. Nie, X. Liu, and X. Liui. An investigation of the effectiveness of concept-based approach in medical information retrieval grium@ clef2014ehealthtask 3. Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2014.

[95] M. S. Simpson, E. Voorhees, and W. Hersh. Overview of the trec 2014 clinical decision support track. In Proc. 23rd Text Retrieval Conference (TREC 2014). National Institute of Standards and Technology (NIST), 2014.

[96] S. Sinha. Integration of prior biological knowledge and epigenetic information enhances the prediction accuracy of the bayesian wnt pathway. Integrative Biology, pages 1034–1048, 2014.

[97] C. A. Smith and P. Z. Stavri. Consumer health vocabulary. In Consumer Health Informatics, pages 122–128. Springer, 2005.

[98] C. A. Sneiderman, D. Demner-Fushman, M. Fiszman, N. C. Ide, and T. C. Rindflesch. Knowledge-based methods to help clinicians find answers in medline. Journal of the American Medical Informatics Association, 14(6):772–780, 2007.

[99] L. Soldaini. The Knowledge and Language Gap in Medical Information Seeking. PhD thesis, Georgetown University, 2018.

[100] L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Query reformulation for clinical decision support search. Technical report, DTIC Document, 2014.

[101] L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Retrieving medical literature for clinical decision support. In Advances in Information Retrieval, pages 538–549. Springer, 2015.

[102] L. Soldaini, A. Yates, and N. Goharian. Learning to reformulate long queries for clinical decision support. Journal of the Association for Information Science and Technology, 68(11):2602–2619, 2017.

[103] P. Sondhi, J. Sun, C. Zhai, R. Sorrentino, and M. S. Kohn. Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. Journal of the American Medical Informatics Association, 19(5):851–858, 2012.

[104] P. Srinivasan. Retrieval feedback in medline. JOURNAL-AMERICAN MEDICAL INFORMATICS ASSOCIATION, 3:157–167, 1996.

[105] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. Citeseer.

[106] A. Sundaram. Information retrieval: A health care perspective. Bulletin of the Medical Library Association, 84(4):591, 1996.

[107] V. Sundararajan, T. Henderson, C. Perry, A. Muggivan, H. Quan, and W. A. Ghali. New icd-10 version of the charlson comorbidity index predicted in-hospital mortality. Journal of clinical epidemiology, 57(12):1288–1294, 2004.

[108] A. Tian and M. Lease. Active learning to maximize accuracy vs. effort in interactive information retrieval. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 145–154. ACM, 2011.

[109] M. Volk, S. Vintar, and P. Buitelaar. Ontologies in cross-language information retrieval. In Wissensmanagement, pages 43–50, 2003.

[110] E. M. Voorhees. Query expansion using lexical-semantic relations. In SIGIR'94, pages 61–69. Springer, 1994.

[111] C. Wang and R. Akella. Concept-based relevance models for medical and semantic information retrieval. In Proceedings of CIKM'15, pages 173–182, 2015.

[112] Y. Wang and H. Fang. Exploring the query expansion methods for concept-based representation. Technical report, DTIC Document, 2014.

[113] Y. Wang, R. Komandur-Elayavilli, M. Rastegar-Mojarad, and H. Liu. Leveraging both structured and unstructured data for precision information retrieval. In Proceedings of TREC 2017, pages 1–17, 2017.

[114] Z. Wang, K. Zhao, H. Wang, X. Meng, and J.-R. Wen. Query understanding through knowledge-based conceptualization. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 2015.

[115] H. Wasserman and J. Wang. An applied evaluation of snomed ct as a clinical vocabulary for the computerized diagnosis and problem list. In AMIA Annual Symposium Proceedings, volume 2003, page 699. American Medical Informatics Association, 2003.

[116] Z. Xie, Y. Xia, and Q. Zhou. Incorporating semantic knowledge with MRF term dependency model in medical document retrieval. In NLPCC'15, pages 219–228. Springer, 2015.

[117] C. Xiong and J. Callan. Esdrank: Connecting query and documents through external semi-structured data. In International Conference on Information and Knowledge Management, volume 6, pages 3–1, 2015.

[118] C. Xiong and J. Callan. Query expansion with freebase. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pages 111–120. ACM, 2015.

[119] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In Proceedings of SIGIR'02, pages 59–66, 2009.

[120] Z. Xu and R. Akella. Active relevance feedback for difficult queries. In Proceedings of the 17th ACM conference on Information and knowledge management, pages 459–468. ACM, 2008.

[121] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 603– 610. ACM, 2008.

[122] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 24th ACM SIGIR, pages 334–342, 2001.

[123] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

[124] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 49–56. ACM, 2002.

[125] C. Zhang and T. Chen. An active learning framework for content-based information retrieval. Multimedia, IEEE Transactions on, 4(2):260–268, 2002.

[126] J. Zhang, L. Lin, S. Diao, Y. Li, R. Liu, W. Xu, and J. Guo. Pris at 2012 trec medical track: Query expansion, retrieval and ranking. Technical report, DTIC Document, 2012.

[127] N. Zhiltsov, A. Kotov, and F. Nikolaev. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 253– 262. ACM, 2015.

[128] M. Zhong and X. Huang. Concept-based biomedical text retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 723–724. ACM, 2006.

[129] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In Proceedings of SIGIR'07, pages 655–662, 2007.

[130] W. Zhu, X. Xu, X. Hu, I.-Y. Song, and R. B. Allen. Using umls-based re-weighting terms as a query expansion strategy. In GrC, pages 217–222, 2006.

ABSTRACT

UTILIZING KNOWLEDGE BASES IN INFORMATION RETRIEVAL FOR CLINICAL DECISION SUPPORT AND PRECISION MEDICINE

by

SAEID BALANESHINKORDAN

May 2019

Advisor: Dr. Alexander Kotov

Major: Computer Science

Degree: Doctor of Philosophy

Accurately answering queries that describe a clinical case and aim at finding articles in a collection of medical literature requires utilizing knowledge bases in capturing many explicit and latent aspects of such queries. Proper representation of these aspects needs knowledge-based query understanding methods that identify the most important query concepts as well as knowledge-based query reformulation methods that add new concepts to a query. In the tasks of Clinical Decision Support (CDS) and Precision Medicine (PM), the query and collection documents may have a complex structure with different components, such as disease and genetic variants that should be transformed to enable an effective information retrieval. In this work, we propose methods for representing domain-specific queries based on weighted concepts of different types whether exist in the query itself or extracted from the knowledge bases and top retrieved documents. Besides, we propose an optimization framework, which allows unifying query analysis and expansion by jointly determining the importance weights for the query and expansion concepts depending on their type and source. We also propose a

probabilistic model to reformulate the query given genetic information in the query and collection documents. We observe significant improvement of retrieval accuracy will be obtained for our proposed methods over state-of-the-art baselines for the tasks of clinical decision support and precision medicine.

AUTOBIOGRAPHICAL STATEMENT

Saeid Balaneshinkordan received his Master's degree in Electrical Engineering at Iran University of Science and Technology in 2012 and his Master's degree in Computer Science at Wayne State University in 2018. He has a bachelors' degree in Electrical Engineering from Urmia University. His major research interests include Information Retrieval, Clinical Decision Support Systems and Deep Learning.