Inferring Conceptual Relationships to Improve Medical Records Search

Nut Limsopatham[1], Craig Macdonald[2], and Iadh Ounis[2]
nutli@dcs.gla.ac.uk[1], firstname.lastname@glasgow.ac.uk[2]

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK

ABSTRACT

Medical records search is challenging because of the inherent implicit knowledge within medical records and queries. Such knowledge is known to the medical practitioners but may be hidden from a search system. For example, when searching for the medical records of patients with a heart disease, medical practitioners commonly know that the medical records of patients taking the amiodarone medicine are relevant, since this drug is used to combat a heart disease. In this paper, we argue that leveraging such implicit knowledge improves the retrieval effectiveness, since it provides new evidence to infer the relevance of medical records towards a query. Specifically, using a novel concept-based representation for both medical records and queries, we expand the queries by inferring additional conceptual relationships from domain-specific resources as well as by extracting informative concepts from the top-ranked medical records. We evaluate the retrieval effectiveness of our proposed approach in the context of the TREC 2011 and 2012 Medical Records track. Our results show the effectiveness of our approach to model the implicit knowledge in medical records search, whereby the infAP retrieval performance is significantly improved up to 14.43% over an effective concept-based representation baseline. Moreover, our proposed approach could achieve retrieval effectiveness comparable to the performance of the best TREC 2011 and 2012 systems.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Experimentation, Performance

Keywords: Medical Records Search, Query and Document Representation, Inference

1. INTRODUCTION

Electronic medical records (EMRs) detail the medical history of patients visiting healthcare providers [15, 27]. Using medical records search systems, these EMRs could be leveraged to aid health practitioners in identifying effective procedures (e.g. diagnostic tests and treatments) for patients visiting a hospital with particular health conditions [11, 12]. However, one of the major challenges of searching in the medical domain is to deal with often complex, inconsistent and ambiguous terminology [16, 19, 28]. For example, it is commonly known by medical practitioners that *'cancer'*, *'carcinoma'*, *'CA'*, and *'malignant tumour'* share a similar meaning. However, such information may be hidden from traditional search systems. To handle such a challenge, prior works resorted to domain-specific resources to improve the representation of medical documents and queries. For instance, by exploiting the knowledge obtained from domain-specific resources, such as MeSH,[1] using concept-based representation approaches, *'cancer'*, *'carcinoma'*, *'CA'*, and *'malignant tumour'* are represented with the same concept to lessen the mismatch of synonymous terms in medical documents and queries [14, 24, 28]. Moreover, synonym and hyponym relationships of medical terms obtained from domain-specific resources have been effectively exploited to reformulate queries in order to further improve retrieval performance (e.g. [5, 25]). Hence, a search system could infer that a medical record of patients suffering from *'pulmonary atresia'* is relevant to a query searching for the medical records of patients suffering from *'a heart disease'*, since according to MeSH, *'pulmonary atresia'* is a particular form of *'a heart disease'*.

Nevertheless, the aforementioned approaches could not leverage the inherent implicit knowledge in medical records and queries, which could be exploited to improve retrieval effectiveness. For example, we can infer that a medical record of a patient is relevant to a query searching for patients with a particular disease, if the patient is treated with a medicine for that disease. Statistical query expansion (QE) approaches, such as pseudo-relevance feedback, indirectly deal with the implicit knowledge challenge by using occurrence statistics of terms in the top-ranked documents to improve the representation of the original query [1]. For example, Limsopatham et al. [17] effectively applied the Divergence from Randomness (DFR) Bo1 QE model [1] to improve the retrieval performance of a medical records search system. However, these QE approaches might not effectively deal with the implicit knowledge challenge, if inherently related concepts are not observed in the top-ranked records.

Instead, to leverage the implicit knowledge inherent to the medical records search process, in this paper we go beyond classical statistical QE techniques. Indeed, we propose to improve the representation of medical queries by inferring the relationships of medical concepts, which are

OAIR'13, May 22-24, 2013, Lisbon, Portugal.

[1] http://www.nlm.nih.gov/mesh/

typically considered by a medical practitioner when dealing with a patient. In particular, we propose to initially represent medical records and queries using only concepts related to *the four aspects of the medical decision criteria (namely, symptom, diagnostic test, diagnosis, and treatment)* [23], as they are important information considered by healthcare practitioners when consulting with a patient. Next, using this new conceptual representation of the medical records and queries, we deploy a novel query expansion approach to further improve the query representation by exploiting two types of resources. Firstly, we use knowledge gained from external resources to infer the relationships between concepts using the four aforementioned aspects. For example, we can infer that a patient suffers from a particular disease, if the patient takes a particular medicine. Secondly, we extract the most informative concepts from the top-ranked medical records. Importantly, the concepts inferred using these two types of resources are used to expand the original queries to improve their representation.

We evaluate our proposed approach in the context of the TREC 2011 [30] and 2012 [29] Medical Records track. Our results attest the effectiveness of our proposed approach, as it can significantly improve the retrieval performance over an effective concept-based representation baseline. We also find that our proposed approach markedly outperforms the system that leverages only either of the two resource types. Moreover, our achieved retrieval effectiveness is comparable to the performance of the best systems at TREC 2011 and 2012 [29, 30].

The main contributions of this paper are threefold:

1. We introduce a task-specific representation approach to effectively represent medical records and queries using medical concepts based on the four aspects of medical decision criteria.
2. We propose a novel query expansion approach that models the relationships between concepts using the four aspects of medical decision criteria, by leveraging medical knowledge gained from external resources and the occurrence statistics of concepts in the top-ranked medical records to improve the query representation and to infer relevance.
3. We thoroughly evaluate our proposed approach within the standard experimentation paradigm provided by the TREC 2011 and 2012 Medical Records track.

The remainder of this paper is organised as follows. In Section 2, the backgrounds of searching medical records and related works are discussed. Section 3 proposes our novel query expansion approach that leverages the conceptual relationships gained from both the external resources and occurrence statistics of concepts in the top-ranked medical records to improve retrieval performance. Experimental setup and results are presented in Sections 4 and 5, respectively. In Section 6, we further evaluate our proposed conceptual QE approach when combined with a term-based representation technique. Finally, we conclude the paper in Section 7.

2. RELATED WORK

Electronic medical records (EMRs), which detail the healthcare information of patients, have been developed to improve the quality of healthcare services [12]. For instance, EMRs could be exploited to identify treatments that have been used effectively to combat a particular disease [11, 12]. However, the characteristics of medical records and queries, such as the complexity of the medical terminology, are different from those of other domains. Hence, effective search approaches for medical records are needed. In 2011, TREC developed a search task to facilitate the research in this area [30]. In particular, the task of the TREC Medical Records track [29, 30] aims to rank patients with respect to the relevance of their medical records towards a query.

Prior work (e.g. [10, 18, 34]) effectively handled this search task using techniques previously developed for expert search [7], since the goal of both tasks is to rank people (i.e. patients or expert persons) based on the relevance of their associated documents. On one hand, expert search aims to rank experts based on the relevance of the documents they have written or that mention them [7]. On the other hand, medical records search ranks patients based on the relevance of their medical records. Hence, in this work, we also handle medical records search using well-established approaches previously developed for expert search, which use ranked medical records to rank patients (e.g. Voting Model [20] and Model 2 [6]). Specifically, the Voting Model ranks patients using a voting process, where the ranking of medical records (denoted $R(Q)$) defines the relevance scores for the patients to be retrieved. Each retrieved medical records in $R(Q)$ is said to vote for the relevance of its associated candidate patient using voting techniques such as, CombMAX, CombSUM, expCombSUM. Indeed, each voting technique firstly ranks medical records based on their relevance towards a query using a traditional weighting model (e.g. BM25 [22], DFR DPH [2]), and then aggregates the votes from the medical records to their associated patients, to create a ranked list of likely relevant patients for the query [20].

One of the important research areas of searching in the medical domain is dealing with the complexity, ambiguity and inconsistency of the medical terminology [16, 19, 28]. For example, when referring to *'coronary heart disease'*, different medical practitioners may use terms, such as *'coronary artery disease'*, *'arteriosclerotic heart disease'*, *'CHD'*, or *'CAD'*. Previous work resorted to domain-specific resources to handle such a challenge [16, 19, 28]. For instance, Srinivasan [24] and Trieschnigg et al. [28] represented medical documents and queries using medical concepts obtained from domain-specific resources, such as MeSH, to alleviate the synonymous mismatch of terms in medical documents and queries. Aronson [3] deployed MetaMap [4] – a medical concept recognition tool based on the UMLS Metathesaurus[2] – to identify all concept, in the medical documents and queries, and to represent them in the form of the UMLS Concept Unique Identifier (CUI). However, concept representation approaches are effective only when combined with a traditional term-based representation [24, 28]. Moreover, in the form of query expansion (QE), synonyms and hyponyms of concepts in the medical documents and queries have also been used to improve the representation of medical queries [5, 25]. For example, Aronson and Rindflesch [5], and Srinivasan [25] effectively expanded concepts in a query with their synonyms and hyponyms obtained from domain-specific resources, such as MeSH and UMLS Metathesaurus. Later, a technique called *concept-based retrieval* was proposed to semantically handle the challenge [26, 32, 33]. It identifies medical concepts in the query and exploits domain-specific resources to expand the query with its associated terms, while ranking. Our work differs from these previous approaches, in that to alleviate the challenge of the complex, ambiguous and inconsistent terminology, we propose

[2] http://www.nlm.nih.gov/research/umls/

to represent medical records and queries by using only concepts that are related to *the four aspects of the medical decision criteria* [23], as they are essential information for health practitioners when dealing with a patient.

In addition, QE techniques, such as pseudo-relevance feedback, have been used to effectively improve the representation of queries in different search tasks [1]. In particular, the approach is to expand a query with *a set of informative terms* obtained from the top-ranked documents. In the context of searching the medical domain, Srinivasan [25] and Limsopatham et al. [17] reported that pseudo-relevance feedback can effectively improve retrieval performance. In particular, pseudo-relevance feedback can indirectly deal with the implicit knowledge, since it may expand the query with semantically-related concepts. For example, for a query searching for *"patients with heart disease"*, the expanded terms might be *'amiodarone'* and *'angina'*, which are a treatment and a symptom associated to *'heart disease'*. Importantly, in medical records such relationships between concepts related to the aspects of the medical decision criteria are strongly established. On the other hand, from a medical record of a patient taking the *'amiodarone'* medicine (treatment), healthcare practitioners can infer that the patient is suffering from *'heart disease'* (diagnosis), since *'amiodarone'* is used to combat *'heart disease'*. However, pseudo-relevance feedback may not always be able to leverage such implicit knowledge, if the associated concepts do not appear in the top-ranked medical records. Hence, in this work, we propose to also directly infer these relationships to improve the query representation. Specifically, we propose a new QE approach that uses conceptual relationships extracted from both external resources and the top-ranked medical records to improve retrieval effectiveness.

3. INFERRING CONCEPTUAL RELATIONSHIPS

In this section, we propose our approach to infer medical conceptual relationships in order to improve the representation of medical queries, and hence enhance retrieval performance. In particular, in Section 3.1, we first discuss our task-specific representation approach to represent medical records and queries in the forms of concepts. Section 3.2 introduces our query expansion approach that models the relationships between concepts associated to the aspects of the medical decision criteria using external resources. Finally, Section 3.3 discusses how the occurrence statistics of concepts in the top-ranked medical records are used to infer conceptual relationships within our query expansion approach.

3.1 Task-Specific Representation

We first introduce our task-specific representation approach. While traditional concept-based representation approaches (e.g. [14, 24]) represent medical records and queries using all identified medical concepts, we propose to represent them by focusing only on concepts related to the four aspects of the *medical decision criteria* (namely, symptoms, diagnostic tests, diagnoses, and treatments), which are derived from the medical decision making process described by Silfen et al. [23]. Indeed, these criteria are considered by medical practitioners when dealing with a patient, including problems (symptoms and diagnoses), diagnostic procedures (diagnostic tests), and management options (treatments). For example, knowing that a patient visiting a hospital with *'chest pain'* (symptom), a healthcare practitioner may suspect that the patient has *'heart disease'* (diagnosis). Hence, given the symptom, the practitioner compiles a set of diagnostic procedures, such as *'chest X-ray'* (diagnostic test) for the patient. Once the practitioner is confident that the patient suffers from *'heart disease'* (diagnosis), the practitioner may prescribe a treatment, such as *'coronary artery bypass surgery'*, for the patient. We hypothesise that these inferences can be effectively used to identify medical records relevant to a query. We deploy MetaMap [4] to extract concepts, in medical records and queries, and represent the identified concepts in the form of the UMLS Concept Unique Identifier (CUI). Importantly, we use only the concepts related to the four aforementioned aspects, which we identify based on their MetaMap's semantic type.[3] In Table 1, we list the 16 MetaMap's semantic types that are associated to the medical decision criteria. Medical concepts with the MetaMap's semantic type defined in the first column are associated to an aspect of the medical decision criteria, if there is a tick (✔) in the column of that aspect. For example, concepts having the MetaMap's semantic type *Disease or Syndrome* are associated with the *diagnosis* aspect. Table 2 shows the concepts obtained from the query *"patients with diabetes mellitus who also have thrombocytosis"*, using our task-specific representation approach. Indeed, our approach identifies two diagnosis concepts: *'diabetes mellitus' (C0011849)* and *'thrombocytosis' (C0836924)*, from the query.

Concept (CUI)	MetaMap's Definition	Related Aspects
C0011849	Diabetes Mellitus [Disease or Syndrome]	Diagnosis
C0836924	Thrombocytosis [Disease or Syndrome]	Diagnosis

Table 2: An example of medical concepts obtained from the query *"patients with diabetes mellitus who also have thrombocytosis"* using our task-specific representation approach.

3.2 Conceptual Association-based QE

Next, we hypothesise that the strongly established relationships between medical concepts related to the medical decision criteria could be leveraged to deal with the implicit knowledge challenge. For example, from evidence that a patient takes *'olmesartan'* medicine (treatment), we can infer that the patient suffers from *'hypertension'* (diagnosis), since *'olmesartan'* is a treatment for *'hypertension'*. Therefore, using external domain-specific resources, we propose to enhance the representation of queries by inferring the relationships between concepts related to the four aspects of the medical decision criteria. Indeed, we reformulate the queries by using association rules extracted from medical resources (e.g. ontologies and health-related websites).

Firstly, driven by the four aspects of the medical decision criteria, we extract directed association rules representing the relationships between concepts from two different types of medical resources, which are ontology-based and free-text-based resources, respectively. We use different strategies for extracting conceptual relationships from each type of the resources. For ontology-based resources (e.g. MedDRA[4] and DOID[5]), we use the semantic relationships of concepts within each ontology to represent the relationships between concepts. For free-text-based resources (e.g. `http://www.rxlist.com`), we use MetaMap to identify concepts from the free-text, and then infer the relationships between

[3] `http://metamap.nlm.nih.gov/SemanticTypeMappings_2011AA.txt`

[4] `http://www.meddramsso.com`

[5] `http://purl.bioontology.org/ontology/DOID`

MetaMap's Semantic Type	Aspects of the Medical Decision Criteria			
	Symptom	Diagnostic test	Diagnosis	Treatment
Body Location or Region	✔	✔	✔	✔
Body Part, Organ, or Organ Component	✔	✔	✔	✔
Clinical Drug	–	–	–	✔
Diagnostic Procedure	–	✔	–	–
Disease or Syndrome	–	–	✔	–
Finding	✔	–	–	–
Health Care Activity	–	✔	–	✔
Injury or Poisoning	✔	–	–	–
Intellectual Product	–	✔	–	✔
Medical Device	–	✔	–	✔
Mental or Behavioral Dysfunction	✔	–	✔	–
Neoplastic Process	✔	✔	✔	✔
Pathologic Function	✔	–	–	–
Pharmacologic Substance	–	–	–	✔
Sign or Symptom	✔	–	–	–
Therapeutic or Preventive Procedure	–	–	–	✔

Table 1: List of 16 of the MetaMap's 133 semantic types that we consider for our proposed approach, based on the four aspects of the medical decision criteria.

the identified concepts. For example, from a drug indication in the `http://www.rxlist.com` website, which states that *"Boniva (ibandronate sodium) is indicated for the treatment and prevention of osteoporosis in postmenopausal women"*, MetaMap can identify concepts *'Boniva'* (treatment) and *'osteoporosis'* (diagnosis). Assuming relationships between medical concepts found in the drug description, we surmise that there is an association between the two concepts. Next, the extracted association rules are stored in a one-to-many relationships database. For instance, as shown in Figure 1, the rules associated to the concept *'osteoporosis'* are *'Dowager's hump'*→*'osteoporosis'*, *'DEXA'*→*'osteoporosis'*, *'Prolia'*→*'osteoporosis'*, and *'Boniva'*→*'osteoporosis'*. These association rules provide new evidence to infer the relevance of medical records. For instance, we can infer that patients taking *'Boniva'* medicine suffer from *'osteoporosis'*, since *'Boniva'* is a treatment for *'osteoporosis'*.

Secondly, during retrieval, we exploit these extracted association rules to reformulate the queries. In particular, we first retrieve a set of candidate concept expansions (denoted $inferred(Q)$) corresponding to the query concepts from the extracted association rules. Then, to prevent excessively general candidate concepts being added to the query, we estimate the association of a query concept and each candidate concept expansion using a Bayesian probabilistic score computed based on the occurrences of concepts in the association rules (both derived from ontologies and free-text resources). The higher the probability, the stronger the relationship between the two concepts. Indeed, the Bayesian probabilistic score of the association between query concept t and its corresponding concept t' is estimated as follows:

$$w_a(t,t') = p(t'|t) = \frac{p(t' \cap t)}{p(t)} \tag{1}$$

where $p(t' \cap t)$ is the maximum likelihood that the concept t' co-occurs with the query concept t within all the extracted association rules, and $p(t)$ is the maximum likelihood that the concept t is contained in an association rule. This is calculated based on the count of the appearance of each concept in the whole association rules database.

Figure 1 shows an example of how our conceptual association-based query expansion (QE) approach identifies candidate concept expansions for the query *"patients with osteoporosis"*. In particular, we first obtain the concept *'osteoporosis'* from the query using our task-specific representation approach. Then, the concept *'osteoporosis'* is used to retrieve related concepts from the database of association rules previously extracted from the medical resources. The retrieved candidate concept expansions include *'Dowager's hump'*, *'DEXA'*, *'Prolia'* and *'Boniva'*, which are the symptom, diagnostic test, and treatments associated to the original query concept (i.e. *'osteoporosis'*). As it has been shown that QE is effective when the highly informative terms are added to the original query [1], we follow Amati [1] and use only the *top 10* candidate concept expansions, which are ranked based on the score computed using Equation (1), to expand the original query.

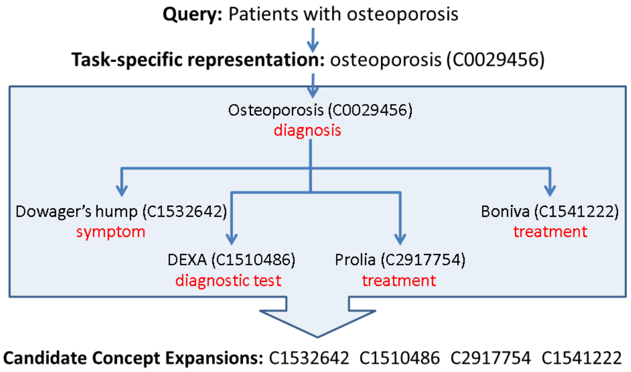
Query: Patients with osteoporosis
Task-specific representation: osteoporosis (C0029456)
Osteoporosis (C0029456)
diagnosis
Dowager's hump (C1532642)
symptom
DEXA (C1510486)
diagnostic test
Prolia (C2917754)
treatment
Boniva (C1541222)
treatment
Candidate Concept Expansions: C1532642 C1510486 C2917754 C1541222

Figure 1: An example of identifying candidate concept expansions using our conceptual association-based approach on query *"patients with osteoporosis"*.

3.3 Local Statistics-based QE

From the medical records and queries represented using our task-specific representation approach, we enhance the representation of the queries by using the information of the top-ranked medical records. To deal with the implicit knowledge challenge, we infer the concepts that are related to a query from the occurrence statistics of medical concepts in the top-ranked medical records. In particular, we apply pseudo-relevance feedback to expand the original queries with the informative concepts extracted from the top-ranked medical records. We select the most informative concepts from the top-ranked medical records retrieved using the original query. Indeed, the concepts occurring in the top-ranked medical records are firstly weighted and ranked using a term weighting model. Then, the top-ranked concepts (i.e. the most informative concepts) are used to expand the original query. For example, for the query *"patients with vascular disease"*, which are represented as *"C0042373"* (*'vascular disease'*), a pseudo-relevance feedback QE approach could identify related concepts, such as *'C0190932'* (*'femoral-popliteal artery bypass graft'*) and *'C0014098'* (*'endarterectomy'*), which are treatments for the disease. However, the related diagnostic procedures and symptoms, such as *'C0202896'* (*'carotid angiogram'*), might not be added to the query, if they do not appear in the top-ranked medical records.

Hence, in this work, we leverage both the external resources and the local statistics from the top-ranked medical records to improve the representation of the queries. Indeed, our proposed QE approach infers relationships of concepts using both the association rules extracted from the domain-specific resources and the occurrence statistics of concepts in the top-ranked medical records. We estimates the relevance of a medical record d towards the query Q as follows:

$$score(d,Q) = \sum_{t'' \in Q_e} qtw(t'') \cdot score_t(d,t'') + \lambda_r \cdot \sum_{t \in Q} \sum_{t' \in inferred(Q)} w_a(t,t') \cdot score_t(d,t') \qquad (2)$$

where t'' and $qtw(t'')$ is a concept and its term weight in the expanded query Q_e, which is reformulated using the occurrence statistics of concepts in the top-ranked medical records using any QE model. λ_r is a parameter to weight the importance of the relevance score of the concepts expanded using the conceptual association-based approach. $score_t()$ can be calculated using any term weighting model such as BM25, $inferred(Q)$ returns a set of the top 10 concepts related to the concepts in the original query Q ranked based on the score $w_a(t,t')$ computed using Equation (1).

4. EXPERIMENTAL SETUP

We have introduced our proposed QE approach to infer conceptual relationships to improve retrieval performance in Section 3. In particular, we hypothesise that our approach to represent medical records and queries (Section 3.1), and expand the queries, by inferring conceptual relationships using the information from the external resources (Section 3.2) and the top-ranked medical records (Section 3.3), could enhance retrieval performance. In this section, we discuss our experimental setup to evaluate our approach.

4.1 Corpus/Queries/Measures

We evaluate our proposed approach in the context of the TREC 2011 [30] and 2012 [29] Medical Records track. In this track, the task is to identify relevant patient *visits* for a given query topic. Each visit contains all of the medical records associated with a patient's visit to a hospital. Due to privacy concerns [30], a *visit* is used to represent a *patient* as a unit of retrieval. The medical records collection consists of approximately 102k medical records, which can be mapped to 17,265 patient visits, from the University of Pittsburgh NLP Repository.[6] We evaluate our proposed approach using the 34 and 47 topics from the TREC 2011 and 2012 Medical Records track, respectively.

We evaluate the retrieval effectiveness of our proposed approach, in terms of bpref measure [8] for TREC 2011, and in terms of infAP and infNDCG measures [31] for TREC 2012. In particular, bpref is the official measure of TREC 2011, since the absolute number of judged visits per topic is relatively small [30]. bpref is designed for evaluating environments with incomplete relevance data and penalises a system which ranks a judged non-relevant document above a judged relevant document [8]. Both infAP and infNDCG are the official measures of the TREC 2012 Medical Records track [29], since the gold standard judgements are incomplete; hence a sampling approach is deployed to infer the MAP and NDCG performance, respectively.

4.2 Medical Records & Visits Ranking

We index the medical records using the Terrier retrieval platform [21], applying Porter's English stemmer and removing stopwords for term-based representation. For both term- and concept-based representations, the effective parameter-free DPH term weighting model [2] is used to rank medical records (e.g. $score_t()$ in Equation (2)). Then, to rank the patient visits, as explained in Section 2, we deploy the expCombSUM voting technique [20], which gives more importance to the highly relevant medical records while voting for the relevance of the patients. In particular, for a given ranking of medical records ($R(Q)$) with respect to query Q, each medical record is said to vote for the relevance of its associated patient visit. The number of medical records in $R(Q)$ to vote for the relevance of the patient visits is limited to 5,000, as suggested in [18].

4.3 Conceptual Association-based QE

To evaluate our proposed QE approach, for the association rules introduced in Section 3.2, we use the external resources listed in Table 3, which are representatives of both ontology-based and free-text-based medical resources. The types of relationships within the aspects of the medical decision criteria that are extracted using each resource are described in the association type column. Table 4 shows the number of association rules between concepts extracted from the domain-specific resources. Note that there may be association rules that overlap between resources. In total, there are 101,133 extracted association rules in our database.

Resources	Association Types	# of Rules
DOID hierarchy	Specific-general	2,046
MeSH	Specific-general	919
MedDRA	Specific-general	84,898
DOID	Diagnosis-symptom	7,680
`http://www.rxlist.com`	Treatment-diagnosis	4,568
`http://www.webmd.com`	Diagnostic test-diagnosis	1,053

Table 4: Number of association rules extracted from each domain-specific resource.

4.4 Local Statistics-based QE

We deploy a parameter-free Bose-Einstein statistics-based (Bo1) model from the Divergence from Randomness (DFR) framework [1] to extract informative concepts from the top-ranked medical records. The Bo1 QE model calculates the weight (i.e. informativeness) of concepts, as follows [1]:

$$w(t) = tf_x \cdot log_2 \frac{1 + P_n(t)}{P_n(t)} + log_2(1 + P_n(t)) \qquad (3)$$

$$P_n(t) = \frac{F(t)}{N} \qquad (4)$$

where tf_x is the frequency of the query concept t in the top-ranked medical records, $F(t)$ is the frequency of concept t in the collection, and N is the number of medical records in the collection. Following Amati [1], we extract the 10 most informative concepts (i.e. concepts having highest $w(t)$ scores)

[6] `http://www.dbmi.pitt.edu/nlpfront`

Resource	Description of Extracted Association Rules
DOID hierarchy	Hierarchical relationships between concepts within the same diagnosis aspect e.g. a general disease and a specific type of the general disease
MeSH	Hierarchical relationships between concepts within the same aspects e.g. a general disease and a specific type of the general disease
MedDRA	Hierarchical relationships between concepts within the same aspects e.g. a general disease and a specific type of the general disease
DOID	Relationships between symptom concepts and diagnosis concepts e.g. a disease and its symptoms
`http://www.rxlist.com`	Relationships between diagnosis concepts and treatment concepts e.g. a medicine and the diseases that it can remedy
`http://www.webmd.com`	Relationships between diagnostic test concepts and diagnosis concepts e.g. a diagnostic test and the diseases that it can diagnosed

Table 3: List of resources used for extracting the conceptual relationships related to the four medical aspects.

from the top 3 retrieved medical records to reformulate the query. Note that the original query concepts may also appear in the 10 extracted concepts.

Then, the query concept weight qtw of each expanded query concepts can be calculated as:

$$qtw(t) = \frac{qtf}{qtf_{max}} + \frac{w(t)}{lim_{F\rightarrow tfx}w(t)} \qquad (5)$$

$$= F_{max} \cdot log_2\frac{1+P_{n,max}}{P_{n,max}} + log_2(1+P_{n,max})$$

$$P_{n,max} = \frac{F_{max}}{N} \qquad (6)$$

where $lim_{F\rightarrow tfx}w(t)$ is the upper bound of $w(t)$, F_{max} is the frequency F of the concept with the maximum $w(t)$ in the top-ranked medical records. If an original query concept t does not appear in the most informative concepts extracted from the top-ranked medical records, its query term weight qtw remains equal to the original one.

5. EXPERIMENTAL RESULTS

This section presents the experimental results conducted using our proposed QE approach. In particular, Section 5.1 compares the effectiveness of our task-specific representation approach introduced in Section 3.1, with traditional term- and concept-based representation baselines. Section 5.2 discusses the retrieval performance of our QE approach that infers relationships of concepts from the external resources and from the occurrence statistics of concepts in the top-ranked medical records.

5.1 Task-Specific Representation

We first evaluate the effectiveness of our task-specific representation approach. We hypothesise that our approach could effectively represent medical records and queries, since it focuses on representing only concepts that are essential for the medical decision process. Hence, we compare the retrieval effectiveness of our approach with the traditional term- and concept-based representation baselines, which represent medical records and queries using all identified terms and concepts, respectively.

Table 5 shows the retrieval performance of our task-specific representation approach compared to the traditional term- and concept-based representation baselines. Significant differences from the concept-based representation baseline according to the paired t-test are denoted * ($p < 0.05$) and ** ($p < 0.01$). From Table 5, we observe that our task-specific representation approach that focuses only on the concepts related to the four aspects of medical decision criteria markedly outperforms both baselines on both TREC 2011 and 2012 test collections. Indeed, our approach significantly outperforms the concept-based representation baseline on all the retrieval performance measures. For TREC 2011, bpref is significantly ($p < 0.01$) increased from 0.4330 to 0.4929 (+13.83%). For TREC 2012, the infNDCG and infAP are also significantly ($p < 0.01$ and $p < 0.05$) increased from 0.3808 to 0.4218 (+10.77%) and from 0.1662 to 0.1920 (+15.52%), respectively.

Representation Approaches	2011	2012	
	bpref	infNDCG	infAP
Term-based	0.4871	0.4167	0.1703
Concept-based	0.4330	0.3808	0.1662
Task-Specific	**0.4929****	**0.4218****	**0.1920***

Table 5: Comparing retrieval performances of different medical records and query representation approaches on TREC 2011 and 2012 Medical Records track's test topics. Statistical significance (paired t-test) at $p < 0.05$ and $p < 0.01$ over the corresponding concept-based representation baseline are denoted * and **, respectively.

5.2 Inferring Conceptual Relationships

Next, we evaluate the effectiveness of our proposed QE approach that builds upon the task-specific representation. Indeed, our QE approach infers conceptual relationships from medical knowledge extracted from the external resources (Section 3.2) and from the occurrence statistics of concepts in the top-ranked medical records (Section 3.3). We compare the retrieval effectiveness of our proposed approach with the baselines, where only either the knowledge from the external resources or the occurrence statistics of medical concepts in the top-ranked medical records is used to enhance the representation of the queries. In addition, the retrieval performance of the task-specific representation approach (i.e. no QE applied) is also reported.

Our proposed QE approach requires a parameter λ_r as per Equation (2) to be properly set. Hence, to have a fair train/test setting, when conducting experiments on TREC 2011 Medical Records track test collection, we set λ_r based on the best infAP retrieval performance achieved on the TREC 2012 test collection, and for the experiments on the TREC 2012 test collection, we set λ_r based on the best bpref retrieval performance attained on the TREC 2011 test collection. We refer to this parameter setting method as *cross-collection setting*. Furthermore, to see how the parameter setting impacts on the retrieval performance and the potential effectiveness of our QE approach, the performances achieved when using the best λ_r for each retrieval measure on each test collection (i.e. best setting) are also reported.

Table 6 shows the retrieval performance of our proposed QE approach on the TREC 2011 and 2012 test collections. We observe that for the TREC 2011 test collection, our proposed QE approach significantly (paired t-test, $p < 0.05$) improves the bpref retrieval performance from 0.4929 to 0.5250 (+6.51%) over the task-specific representation baseline where no QE is applied. For TREC 2012 test collection, our proposed QE approach also significantly outperforms the task-specific representation baseline, in terms of both infNDCG ($p < 0.05$) and infAP ($p < 0.01$) measures. In particular, the infNDCG retrieval performance is improved from 0.4218 to 0.4534 (+7.49%) and infAP is increased from 0.1920 to 0.2128 (+10.83%). Additionally, as expected, we find that with a proper setting of the parameter λ_r (i.e. the λ_r that results in the best retrieval performance for each topic set), our proposed QE approach can achieve a better retrieval performance.[7] In particular, the infNDCG retrieval performance is improved to 0.4745 (+12.49% over the task-specific representation baseline). This performance significantly outperforms the task-specific representation, local statistics-based QE, and our proposed QE approaches (cross-collection setting), at $p < 0.001$, $p < 0.05$, and $p < 0.05$, respectively. In term of infAP retrieval performance, with the best λ_r, our proposed approach markedly outperforms the task-specific representation baseline (14.43% improvement from infAP 0.1920 to 0.2197). Moreover, we find that our QE approach when using both external resources and information from the top-ranked medical records markedly outperform when using only either of them.

6. RELEVANCE SCORE COMBINATION

Finally, as previous work (e.g. [24]) showed that a concept-based representation approach is effective only when combined with a term-based representation approach, we further evaluate the effectiveness of our proposed QE approach when combined with a term-based representation approach. In particular, we follow the approach by Srinivasan [24], which we refer to as the *relevance score combination* approach, to linearly combine the relevance scores calculated using both the term-based representation approach and our proposed approach (cross-collection setting), as follows [24]:

$$\begin{aligned} score(d,Q) =&\delta \cdot score_{term-based}(d,Q) \\ &+ score_{our-approach}(d,Q) \end{aligned} \tag{7}$$

where δ is a parameter to emphasise the relevance score computed using the term-based representation, which is set to 2.00, as suggested in [24].

In order to have a strong baseline for the term-based representation approach, we follow Diaz and Metzler [9] to improve the representation of the queries by using information from different corpora (we call this the *enriched term-based representation*). In particular, we apply the Bo1 QE model to expand the queries with the top 10 informative terms from the top 3 ranked documents retrieved from the TREC Medical Records and TREC 2005 Genomics [13] track collections, respectively. In addition, we follow [17] to deal with negated language in medical records.

We hypothesise that when combined with the enriched term-based representation approach, our proposed approach could further improve the retrieval effectiveness. In particular, we expect that the retrieval performance of the relevance score combination approach could be better than that of either the enriched term-based representation or our proposed QE approaches. Table 7 compares the retrieval performance of the relevance score combination approach to these two baselines. Moreover, the retrieval performance of the top 3 best TREC 2011 and 2012 Medical Records track systems are also reported.

From Table 7, we observe that the relevance score combination approach markedly outperforms both our proposed QE approach and the enriched term-based representation approach. In particular, for the TREC 2011 Medical Records track test collection, the achieved bpref retrieval performance (0.5764) is markedly better than the TREC best system (0.5520). Furthermore, for the TREC 2012 test collection, in terms of infNDCG retrieval performance, the relevance score combination approach (0.5266) could markedly outperform both the enriched term-based representation (0.4865) and our proposed QE (0.4534) approaches. Specifically, the infNDCG of the relevance score combination is significantly (paired t-test, $p < 0.05$) better than that of the enriched term-based representation, improving by 7.63%. Moreover, the infAP retrieval performance of the relevance score combination is also markedly better than the performance of the two baselines. Indeed, the infAP of the relevance score combination (0.2442) is 12.68% better than the performance of the enriched term-based representation. Hence, our results show that our proposed QE approach could improve the representation and bring novel evidence for a search system to infer the relevance of medical records. Importantly, when combined with a term-based representation (as suggested by [24]), without requiring any training for δ in Equation (7), our approach could further improve the retrieval performance and the achieved performance is comparable to the best systems reported at TREC 2011 and 2012 Medical Records track.

7. CONCLUSIONS

We have proposed our QE approach that infers the relationships between medical concepts to handle the implicit knowledge challenge. In particular, we represent only concepts related to the four aspects of the medical decision criteria, which are essential information that medical practitioners take into account when dealing with patients. Then, our QE approach models relationships between concepts obtained from both domain-specific resources, such as, ontologies and health-related websites, and the local statistics of top-ranked medical records to reformulate the queries. Our results show that our proposed QE approach could effectively enhance the representation of the queries, as it could significantly improve the retrieval performance over an effective concept-based representation baseline up to 14.43%. Furthermore, we showed that our proposed approach could work effectively with a term-based representation approach, since when combining the relevance scores computed using both approaches, the retrieval performance is markedly increased. In particular, the achieved retrieval performance is comparable to the best systems at the TREC 2011 and 2012 Medical Records track, who also deployed other additional techniques (e.g. document segmentation, and document clustering).

For future work, we plan to investigate machine learning techniques to learn to properly set the λ_r parameter of on a per-query basis, and to learn to apply only effective association rules based on the importance and the types of medical concepts to improve the representation of the queries.

8. REFERENCES

[1] G. Amati. Probabilistic Models for Information Retrieval based on Divergence from Randomness. *PhD thesis.* University of Glasgow, 2003.

[2] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track. In *Proc. of TREC*, 2007.

[7] Nevertheless, our training using the fair cross-collection setting achieves a retrieval performance comparable to the best setting.

Approaches	2011	2012	
	bpref	infNDCG	infAP
Task-specific representation	0.4929	0.4218	0.1920
Local Statistics-based QE	**0.5282*** (+7.16%)	0.4530* (+7.49%)	0.2126** (+10.73%)
Conceptual association-based QE (cross-collection setting)	0.4929 (+0%)	0.4198 (-0.47%)	0.1913 (-0.37%)
Conceptual association-based QE (best setting)	0.4929 (+0%)	0.4358 (+3.32%)	0.1981 (+3.2%)
Our proposed QE approach (cross-collection setting)	0.5250* (+6.51%)	**0.4534*** (+7.49%)	**0.2128**** (+10.83%)
Our proposed QE approach (best setting)	0.5283* (+7.18%)	0.4745^{***+o} (+12.49%)	0.2197** (+14.43%)

Table 6: Comparing the retrieval performances of different query representation approaches on TREC 2011 and 2012 Medical Records track's test topics. Statistical significance (paired t-test) at $p < 0.05$, at $p < 0.01$, and at $p < 0.001$ over the task-specific representation baseline are denoted *, ** and ***, respectively. Statistical significance (paired t-test) at $p < 0.05$ over the local statistics-based QE and our proposed QE (cross-collection setting) are denoted $^{+}$ and ^{o}.

2011		2012		
Approaches	bpref	Approaches	infNDCG	infAP
Enriched term-based representation	0.5733	Enriched term-based representation	0.4865	0.2132
Our proposed QE approach	0.5250	Our proposed QE approach	0.4534	0.2128
Relevance score combination	**0.5764**$^{+}$ (+0.54%,+8.92%)	Relevance score combination	**0.5266**$^{*++}$ (+7.63%,+13.91%)	**0.2442** (+12.68%,+12.85%)
Best TREC systems				
CengageM11R3	**0.5520**	udelSUM	**0.5780**	**0.2860**
SCIAMED7	**0.5520**	sennamed2	0.5470	0.2750
UTDHLTCIR	0.5450	atigeo1	0.5240	0.2240

Table 7: Comparing the retrieval performances of different retrieval approaches on TREC 2011 and 2012 Medical Records track's test topics. Statistical significance (paired t-test) at $p < 0.05$, and at $p < 0.01$ over the enriched term-based representation are denoted *, and **. Statistical significance (paired t-test) at $p < 0.05$, and at $p < 0.01$ over our proposed QE approach are denoted $^{+}$, and $^{++}$, respectively.

[3] A. R. Aronson. Exploiting a large thesaurus for information retrieval. In *Proc. of RIAO*, 1994.

[4] A. R. Aronson and F. Lang. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3):229–236, 2010.

[5] A. R. Aronson and T. C. Rindflesch. Query expansion using the UMLS Metathesaurus. In *Proc. of AMIA*, 1997.

[6] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. of SIGIR*, 2006.

[7] K. Balog, P. Thomas, N. Craswell, I. Soboroff, and P. Bailey. Overview of the TREC 2008 Enterprise Track. In *Proc. of TREC*, 2008.

[8] C. Buckley and E. Voorhees. Retrieval Evaluation with Incomplete Information. In *Proc. of SIGIR*, 2004.

[9] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proc. of SIGIR*, 2006.

[10] H. Gurulingappa, B. Uller, M. Hofmann-Apitius, and J. Fluck. A Semantic Platform for Information Retrieval from E-Health Records. In *Proc. of TREC*, 2011.

[11] W. Hersh. Health care information technology: progress and barriers. *J. Am. Med. Assoc.*, 292(18):2273–2274, 2004.

[12] W. Hersh. Information retrieval: A health and biomedical perspective (3rd ed.). *New York : Springer*, 2009.

[13] W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, and M. Hearst. TREC 2005 Genomics Track Overview. In *Proc. of TREC*, 2005.

[14] W. Hersh, D. Hickam, R. Haynes, and K. McKibbon. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *J. Am. Med. Inform. Assoc.*, 1(1):51–60, 1994.

[15] I. Kotsiopoulos, J. Keane, M. Turner, P. Layzell, and F. Zhu. IBHIS: Integration broker for heterogeneous information sources. In *Proc. of AICSAC*, 2003.

[16] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *J. Biomed. Inform.*, 37(6):512–526, 2004.

[17] N. Limsopatham, C. Macdonald, R. McCreadie, and I. Ounis. Exploiting Term Dependence while Handling Negation in Medical Search. In *Proc. of SIGIR*, 2012.

[18] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M. Bouamrane. University of Glasgow at Medical Records track 2011: Experiments with Terrier. In *Proc. of TREC*, 2011.

[19] N. Limsopatham, R. L. T. Santos, C. Macdonald, and I. Ounis. Disambiguating biomedical acronyms using EMIM. In *Proc. of SIGIR*, 2011.

[20] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. of CIKM*, 2006.

[21] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of OSIR at SIGIR*, 2006.

[22] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of SIGIR*, 1994.

[23] E. Silfen. Documentation and coding of ED patient encounters: an evaluation of the accuracy of an electronic medical record. *Am. J. Emerg. Med.*, 24(6):664–678, 2006.

[24] P. Srinivasan. Optimal document-indexing vocabulary for MEDLINE. *Inf. Process. Manage.*, 32(5):503–514, 1996.

[25] P. Srinivasan. Query expansion and medline. *Inf. Process. Manage.*, 32(4):431–443, 1996.

[26] N. Stokes, Y. Li, L. Cavedon, and J. Zobel. Exploring criteria for successful query expansion in the genomic domain. *Inf. Retr.*, 12(1):17–50, 2009.

[27] E. Tambouris, M. H. Willimas, and C. Makropoulos. Co-operative health information networks in Europe: experience from Greece and Scotland. *J. Med. Internet Res.*, 2(2):e11, 2000.

[28] D. Trieschnigg, D. Hiemstra, F. de Jong, and W. Kraaij. A cross-lingual framework for monolingual biomedical information retrieval. In *Proc. of CIKM*, 2010.

[29] E. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *Proc. of TREC*, 2012.

[30] E. Voorhees and R. Tong. Overview of the TREC 2011 Medical Records Track. In *Proc. of TREC*, 2011.

[31] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. of SIGIR*, 2006.

[32] M. Zhong and X. Huang. Concept-based biomedical text retrieval. In *Proc. of SIGIR*, 2006.

[33] W. Zhou, C. Yu, and W. Meng. A system for finding biological entities that satisfy certain conditions from texts. In *Proc. of CIKM*, 2008.

[34] D. Zhu and B. Carterette. Combining Multi-level Evidence for Medical Record Retrieval. In *Proc. of SHB*, 2012.