A Study of Concept-based Weighting Regularization for Medical Records Search

Yue Wang, Xitong Liu, Hui Fang
Department of Electrical & Computer Engineering,
University of Delaware, USA
{wangyue,xtliu,hfang}@udel.edu

Abstract

An important search task in the biomedical domain is to find medical records of patients who are qualified for a clinical trial. One commonly used approach is to apply NLP tools to map terms from queries and documents to concepts and then compute the relevance scores based on the concept-based representation. However, the mapping results are not perfect, and none of previous work studied how to deal with them in the retrieval process. In this paper, we focus on addressing the limitations caused by the imperfect mapping results and study how to further improve the retrieval performance of the concept-based ranking methods. In particular, we apply axiomatic approaches and propose two weighting regularization methods that adjust the weighting based on the relations among the concepts. Experimental results show that the proposed methods are effective to improve the retrieval performance, and their performances are comparable to other top-performing systems in the TREC Medical Records Track.

1 Introduction

With the increasing use of electronic health records, it becomes urgent to leverage this rich information resource about patients' health conditions to transform research in health and medicine. As an example, when developing a cohort for a clinical trial, researchers need to identify patients matching a set of clinical criteria based on their medical records during their hospital visits (Safran et al., 2007; Friedman et al., 2010). This selection process is clearly a domain-specific retrieval problem, which searches for *relevant medical records* that contain useful information about their corresponding patients' qualification to the criteria specified in a query, e.g., "female patient with breast cancer with mastectomies during admission".

Intuitively, to better solve this domain-specific retrieval problem, we need to understand the requirements specified in a query and identify the documents satisfying these requirements based on their semantic meanings. In the past decades, significant efforts have been put on constructing biomedical knowledge bases (Aronson and Lang, 2010; Lipscomb, 2000; Corporation, 1999) and developing natural language processing (NLP) tools, such as MetaMap, to utilize the information from the knowledge bases (Aronson, 2001; McInnes et al., 2009). These efforts make it possible to map free text to concepts and use these concepts to represent queries and documents.

Indeed, *concept-based representation* is one of the commonly used approaches that leverage knowledge bases to improve the retrieval performance (Limsopatham et al., 2013d; Limsopatham et al., 2013b). The basic idea is to represent both queries and documents as "bags of concepts", where the concepts are identified based on the information from the knowledge bases. This method has been shown to be more effective than traditional term-based representation in the medical record retrieval because of its ability to handle the ambiguity in the medical terminology. However, this method also suffers the limitation that its effectiveness depends on the accuracy of the concept mapping results. As a result, directly applying existing weighting strategies might lead to non-optimal retrieval performance.

In this paper, to address the limitation caused by the inaccurate concept mapping results, we propose to regularize the weighting strategies in the concept-based representation methods. Specifically, by applying the axiomatic approaches (Fang and Zhai, 2005), we analyze the retrieval func-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 603–612,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

tions with concept-based representation and find that they may violate some reasonable retrieval constraints. We then propose two concept-based weighting regularization methods so that the regularized retrieval functions would satisfy the retrieval constraints and achieve better retrieval performance. Experimental results over two TREC collections show that both proposed concept-based weighting regularization methods can improve the retrieval performance, and their performance is comparable with the best systems of the TREC Medical Records tracks (Voorhees and Tong, 2011; Voorhees and Hersh, 2012).

Many NLP techniques have been developed to understand the semantic meaning of textual information, and are often applied to improve the search accuracy. However, due to the inherent ambiguity of natural languages, the results of NLP tools are not perfect. One of our contributions is to present a general methodology that can be used to adjust existing IR techniques based on the inaccurate NLP results.

2 Related Work

The Medical Records track of the Text REtrieval Conference (TREC) provides a common platform to study the medical records retrieval problem and evaluate the proposed methods (Voorhees and Tong, 2011; Voorhees and Hersh, 2012).

Concept-based representation has been studied for the medical record retrieval problem (Limsopatham et al., 2013d; Limsopatham et al., 2013b; Limsopatham et al., 2013a; Qi and Laquerre, 2012; Koopman et al., 2011; Koopman et al., 2012). For example, Qi and Laquerre used MetaMap to generate the concept-based representation and then apply a vector space retrieval model for ranking, and their results are one of the top ranked runs in the TREC 2012 Medical Records track (Qi and Laquerre, 2012). To further improve the performance, Limsopatham et al. proposed a task-specific representation, i.e., using only four types of concepts (symptom, diagnostic test, diagnosis and treatment) in the concept-based representation and a query expansion method based on the relationships among the medical concepts (Limsopatham et al., 2013d; Limsopatham et al., 2013a). Moreover, they also proposed a learning approach to combine both term-based and concept-based representation to further improve the performance (Limsopatham et al., 2013b).

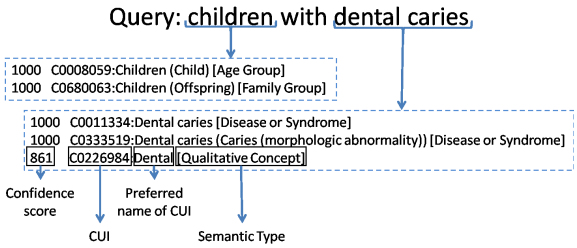

Figure 1: Example of MetaMap result for a query.

Our work is also related to domain-specific IR (Yan et al., 2011; Lin and Demner-Fushman, 2006; Zhou et al., 2007). For example, Yan et al. proposed a granularity-based document ranking model that utilizes ontologies to identify document concepts. However, none of the previous work has studied how to regularize the weight of concepts based on their relations.

It is well known that the effectiveness of a retrieval function is closely related to the weighting strategies (Fang and Zhai, 2005; Singhal et al., 1996). Various term weighting strategies have been proposed and studied for the term-based representation (Amati and Van Rijsbergen, 2002; Singhal et al., 1996; Robertson et al., 1996). However, existing studies on concept-based representation still used weighting strategies developed for term-based representation such as vector space models (Qi and Laquerre, 2012) and divergence from randomness (DFR) (Limsopatham et al., 2013a) and did not take the inaccurate concept mapping results into consideration. Compared with previous work, we focus on addressing the limitation caused by the inaccurate concept mapping. Note that our efforts are orthogonal to existing work, and it is expected to bring additional improvement to the retrieval performance.

3 Concept-based Representation for Medical Records Retrieval

3.1 Problem Formulation

We follow the problem setup used in the TREC medical record track (Voorhees and Tong, 2011; Voorhees and Hersh, 2012). The task is to retrieve relevant patient visits with respect to a query. Since each visit can be associated with multiple medical records, the relevance of a visit is related to the relevance of individual associated medical records. Existing studies computed the relevance

scores at either visit-level, where all the medical records of a visit are merged into a visit document (Demner-Fushman et al., 2012; Limsopatham et al., 2013c), or record-level, where we can first compute the relevance score of individual records and then aggregate their scores as the relevance score of a visit (Limsopatham et al., 2013c; Zhu and Carterette, 2012; Limsopatham et al., 2013d). In this paper, we focus on the visit-level relevance because of its simplicity. In particular, given a patient's visit, all the medical records generated from this visit are merged as a document. Note that our proposed concept-weighting strategies can also be easily applied to record-level relevance modeling.

Since the goal is to retrieve medical records of patients that satisfying requirements specified in a query, the relevance of medical records should be modeled based on how well they match all the requirements (i.e., aspects) specified in the queries.

3.2 Background: UMLS and MetaMap

Unified Medical Language System (UMLS) is a metathesaurus containing information from more than 100 controlled medical terminologies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) and Medical Subject Headings (MeSH). Specifically, it contains the information about over 2.8 million biomedical concepts. Each concept is labeled with a Concept Unique Identifier (CUI) and has a preferred name and a semantic type.

Moreover, NLP tools for utilizing the information from UMLS have been developed. In particular, MetaMap (Aronson, 2001) can take a text string as the input, segment it into phrases, and then map each phrase to *multiple* UMLS CUIs with confidence scores. The confidence score is an indicator of the quality of the phrase-to-concept mapping by MetaMap. It is computed by four metrics: centrality, variation, coverage and cohesiveness (Aronson, 2001). These four measures try to evaluate the mapping from different angles, such as the involvement of the central part, the distance of the concept to the original phrase, and how well the concept matches the phrase. The maximum confidence in MetaMap is 1000.

Figure 1 shows the MetaMap results for an example query "children with dental caries". Two *query aspects*, i.e., "children" and "dental caries", are identified. Each of them is mapped to multiple concepts, and each concept is associated with the confidence score as well as more detailed information about this concept.

3.3 Concept-based Representation

Traditional retrieval models are based on "bag of terms" representation. One limitation of this representation is that relevance scores are computed based on the matching of terms rather than the meanings. As a result, the system may fail to retrieve the relevant documents that do not contain any query terms.

To overcome this limitation, concept-based representation has been proposed to bridge the vocabulary gap between documents and queries (Qi and Laquerre, 2012; Limsopatham et al., 2013b; Koopman et al., 2012). In particular, MetaMap is used to map terms from queries and documents (e.g., medical records) to the semantic concepts from biomedical knowledge bases such as UMLS. Within the concept-based representation, the query can then be represented as a bag of all the generated CUIs in the MetaMap results. For example, the query from Figure 1 can be represented as $\{C0008059, C0680063, C0011334, C0333519, C0226984\}$. Documents can be represented in a similar way.

After converting both queries and documents to concept-based representations using MetaMap, previous work applied existing retrieval functions such as vector space models (Singhal et al., 1996) to rank the documents. Note that when referring to existing retrieval functions in the paper, they include traditional keyword matching based functions such as pivoted normalization (Singhal et al., 1996), Okapi (Robertson et al., 1996), Dirichlet prior (Zhai and Lafferty, 2001) and basic axiomatic functions (Fang and Zhai, 2005).

4 Weighting Strategies for Concept-based Representation

4.1 Motivation

Although existing retrieval functions can be directly applied to concept-based representation, they may lead to non-optimal performance. This is mainly caused by the fact that MetaMap may generate more than one mapped concepts for an aspect, i.e., a semantic unit in the text.

Ideally, an aspect will be mapped to only *one* concept, and different concepts would represent different semantic meanings. Under such a situ-

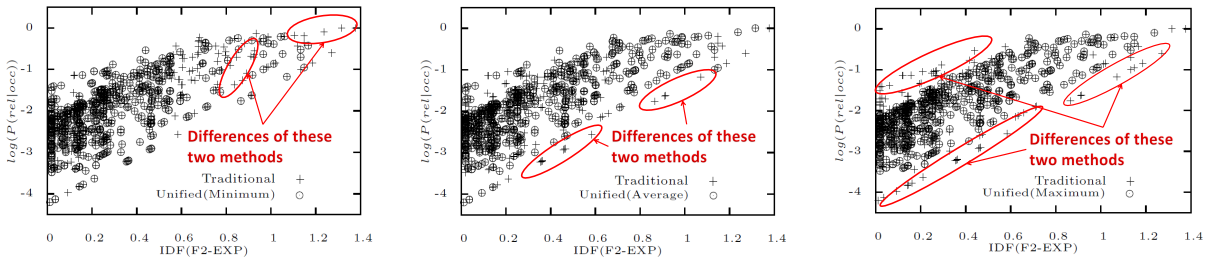

Figure 2: Exploratory data analysis (From left to right are choosing minimum, average and maximum IDF concepts as the representing concepts, respectively. The removed concepts are highlighted in the figures.).

ation, traditional retrieval functions would likely work well and generate satisfying retrieval performance since the relations among concepts are independent which is consistent with the assumptions made in traditional IR (Manning et al., 2008).

However, the mapping results generated by MetaMap are not perfect. Although MetaMap is able to rank all the candidate concepts with the confidence score and pick the most likely one, the accuracy is not very high. In particular, our preliminary results show that turning on the disambiguation functionality provided by MetaMap (i.e., returning only the most likely concept for each query) could lead to worse retrieval performance than using all the candidate mappings. Thus, we use the one-to-many mapping results generated by MetaMap, in which each aspect can be mapped to multiple concepts.

Unfortunately, such one-to-many concept mappings could hinder the retrieval performance in the following two ways.

- The multiple concepts generated from the same aspect are related, which is inconsistent with the independence assumption made in the existing retrieval functions (Manning et al., 2008). For example, as shown in Figure 1, "dental caries" is mapped to three concepts. It is clear that the concepts are related, but existing retrieval functions are unable to capture their relations and would compute the weight of each concept independently.

- The one-to-many mapping results generated by MetaMap could arbitrarily inflate the weights of some query aspects. For example, as shown in Figure 1, query aspect "children" is mapped to 2 concepts while "dental caries" is mapped to 3 concepts. In the existing retrieval functions, term occurrences are important relevance signals. However, when converting the text to concepts representation using MetaMap, the occurrences of the concepts are determined by not only the original term occurrences, a good indicator of relevance, but also the number of mapped concepts, which is determined by MetaMap and has nothing to do with the relevance status. As a result, the occurrences of concepts might not be a very accurate indicator of importance of the corresponding query aspect.

To address the limitations caused by the inaccurate mapping results, we propose to apply axiomatic approaches (Fang and Zhai, 2005) to regularize the weighting strategies for concept-based representation methods. In particular, we first formalize retrieval constraints that any reasonable concept-based representation methods should satisfy and then discuss how to regularize the existing weighting strategies to satisfy the constraints and improve the retrieval performance.

We first explain the notations used in this section. Q and D denote a query and a document with the concept-based representation. $S(Q, D)$ is the relevance score of D with respect to Q. e_i denotes a concept, and $\mathcal{A}(e)$ denotes the query aspect associated with e, i.e., a set of concepts that are mapped to the same phrases as e by using MetaMap. $i(e)$ is the normalized confidence score of the mapping for concept e generated by MetaMap. $c(e, D)$ denotes the occurrences of concept e in document D, $df(e)$ denotes the number of documents containing e. $|D|$ is the document length of D. $Imp_c(e)$ is the importance of the concept such as the concept IDF value, and $Imp_A(\mathcal{A})$ is the importance of the aspect.

4.2 Unified concept weighting regularization

We now discuss how to address the first challenge, i.e,. how to regularize the weighting strategy so that we can take into consideration the fact that concepts associated with the same query aspect are not independent. We call a concept is a *variant* of another one if both of them are associated with the same aspect.

Intuitively, given a query with two aspects, a document covering both aspects should be ranked higher than those covering only one aspect. We can formalize the intuition in the concept-based representation as the following constraint.

Unified Constraint: Let query be $Q = \{e_1, e_2, e_3\}$, and we know that e_2 is a variant of e_3. Assume we have two documents D_1 and D_2 with the same document length, i.e., $|D_1| = |D_2|$. If we know that $c(e_1, D_1) = c(e_3, D_2) > 0$, $c(e_1, D_2) = c(e_3, D_1) = 0$ and $c(e_2, D_1) = c(e_2, D_2) > 0$, then $S(Q, D_1) > S(Q, D_2)$.

It is clear that existing retrieval functions would violate this constraint since they ignore the relations among concepts.

One simple strategy to fix this problem is to merge all the concept variants as a single concept and select one representative concept to replace all occurrences of other variants in both queries and documents. By merging the concepts together, we are aiming to purify the concepts and make the similar concepts centralized so that the assumption that all the concepts are independent would hold.

Formally, the adjusted occurrences of a concept e in a document D is shown as follows:

$$c_{mod}(e, D) = \begin{cases} \sum_{e' \in EC(e)} c(e', D) & e = Rep(EC(e)) \\ 0 & e \neq Rep(EC(e)) \end{cases} \quad (1)$$

where $c(e, D)$ is the original occurrence of concept e in document D, $EC(e)$ denotes a set of all the variants of e including itself (i.e., all the concepts with the same preferred name as e), and $Rep(EC(e))$ denotes the representative concept from $EC(e)$.

It is trivial to prove that, with such changes, existing retrieval functions would satisfy the above constraint since the constraint implies TFC2 constraint defined in the previous study (Fang et al., 2004).

Now the remaining question is how to select the representative concept from all the variants. There are three options: select the concept with the maximum IDF, average IDF, or minimum IDF. We conduct exploratory data analysis on these three options. In particular, for each option, we generate a plot indicating the correlation between the IDF value of a concept and the relevance probability of the concept (i.e., the probability that a document containing the concept is relevant). Note that both original and replaced IDF values are shown in the plot for each option. Figure 2 shows the results. It is clear that the right plot (i.e., selecting the concept with the maximum IDF as the representative concept) is the best choice since the changes make the points less scattered. In fact, this can also be confirmed by experimental results as reported in Table 5. Thus, we use the concept with the maximum IDF value as the representative concept of all the variants.

4.3 Balanced concept weighting regularization

We now discuss how to address the second challenge, i.e., how to regularize the weighting strategy to deal with the arbitrarily inflated statistics caused by the one-to-many mappings.

The arbitrary inflation could impact the importance of the query aspects. For example, as shown in Figure 1, one aspect is mapped to two concepts while the other is mapped to three. Moreover, it could also impact the accuracy of the concept IDF values. Consider "colonoscopies" and "adult", it is clear that the first term is more important than the second one, which is consistent with their term IDF values, i.e., 7.52 and 2.92, respectively. However, with the concept-based representation, the IDF value of the concept "colonoscopies"(C0009378) is 2.72, which is even smaller than that of concept "adult" (C1706450), i.e., 2.92.

To fix the negative impact on query aspects, we could leverage the findings in the previous study (Zheng and Fang, 2010) and regularize the weighting strategy based on the length of query aspects to favor documents covering more query aspects. Since each concept mapping is associated with a confidence score, we can incorporate them into the regularization function as follows:

$$f(e, Q) = (1 - \alpha) + \alpha \cdot \Big(\frac{\sum_{e' \in Q} i(e')}{\sum_{e'' \in \mathcal{A}(e)} i(e'')} \Big), \quad (2)$$

where $i(e)$ is the normalized confidence score of concept e generated by MetaMap, and α is a parameter between 0 and 1 to control the effect of the regularization. When α is set to 0, there is no regularization. This regularization function aims to

penalize the weight of concept e based on its variants as well as the concepts from other aspects. In particular, a concept would receive more penalty (i.e., its weight will be decreased more) when it has more variants and the mappings of these variants are more accurate.

To fix the negative impact on the concept IDF values, we propose to regularize the weighting based on the importance of the query aspect. This regularization can be formalized as the following constraint.

Balanced Constraint: Let Q be a query with two concepts and the concepts are associated with different aspects, i.e., $Q = \{e_1, e_2\}$, and $\mathcal{A}(e_1) \neq \mathcal{A}(e_2)$. Assume D_1 and D_2 are two documents with the same length, i.e., $|D_1| = |D_2|$, and they cover different concepts with the same occurrences, i.e., $c(e_1, D_1) = c(e_2, D_2) > 0$ and $c(e_2, D_1) = c(e_1, D_2) = 0$. If we know $Imp_c(e_1) = Imp_c(e_2)$ and $Imp_A(\mathcal{A}(e_1)) < Imp_A(\mathcal{A}(e_2))$, then we have $S(Q, D_1) < S(Q, D_2)$.

This constraint requires that the relevance score of a document should be affected by not only the importance of the concepts but also the importance of the associated query aspect. In a way, the constraint aims to counteract the arbitrary statistics inflation caused by MetaMap results and balance the weight among concepts based on the importance of the associated query aspects. And it is not difficult to show that existing retrieval functions violate this constraint.

Now the question is how to revise the retrieval functions to make them satisfy this constraint. We propose to incorporate the importance of query aspect into the previous regularization function in Equation (2) as follows:

$$f(e, Q) = (1-\alpha) + \alpha \cdot \left(\frac{\sum_{e' \in Q} i(e')}{\sum_{e'' \in \mathcal{A}(e)} i(e'')} \right) \cdot Imp_A(\mathcal{A}(e)). \quad (3)$$

Note that $Imp_A(\mathcal{A}(e))$ is the importance of a query aspect and can be estimated based on the terms from the query aspect. In this paper, we use the maximum term IDF value from the aspect to estimate the importance, which performs better than using minimum and average IDF values as shown in the experiments (i.e., Table 6). We plan to study other options in the future work.

4.4 Discussions

Both proposed regularization methods can be combined with any existing retrieval functions. In this paper, we focus on one of the state of the art weighting strategies, i.e., F2-EXP function derived from axiomatic retrieval model (Fang and Zhai, 2005), and explain how to incorporate the regularization methods into the function.

The original F2-EXP retrieval function is shown as follows:

$$S(Q, D) = \sum_{e \in Q \cap D} c(e, Q) \cdot (\frac{N}{df(e)})^{0.35} \cdot \frac{c(e, D)}{c(e, D) + b + \frac{b \times |D|}{avdl}} \quad (4)$$

where b is a parameter control the weight of the document length normalization.

With the unified concept weighting regularization, the revised function based on F2-EXP function, i.e., *Unified*, is shown as follows:

$$S(Q, D) = \sum_{e \in Q \cap D} c_{mod}(e, Q) \cdot (\frac{N}{df(t)})^{0.35} \cdot \frac{c_{mod}(e, D)}{c_{mod}(e, D) + b + \frac{b \times |D|}{avdl}} \quad (5)$$

where $c_{mod}(e, D)$ and $c_{mod}(e, Q)$ denote the modified occurrences as shown in Equation (1). It can be shown that this function satisfies the unified constraint but violates the balanced constraint.

Following the similar strategy used in the previous study (Zheng and Fang, 2010), we can further incorporate the regularization function proposed in Equation (3) to the above function to make it satisfy the balanced constraint as follows:

$$S(Q, D) = \sum_{e \in Q \cap D} c_{mod}(e, Q) \cdot (\frac{N}{df(t)})^{0.35} \cdot f(e, Q) \cdot \frac{c_{mod}(e, D)}{c_{mod}(e, D) + b + \frac{b \times |D|}{avdl}} \quad (6)$$

where $f(e, Q)$ is the newly proposed regularization function as shown in Equation (3). This method is denoted as *Balanced*, and can be shown that it satisfies both constraints.

Table 1: Statistics of collections.

	# of unique tokens	AvgDL	AvgQL11	AvgQL12
Term	263,356	2,659	10.23	8.82
Concept	58,192	2,673	8.79	7.81

5 Experiments

5.1 Experiment Setup

We conduct experiments using two data sets from the TREC Medical Records track 2011 and 2012.

Table 2: Description of Methods

Name	Representation	Ranking strategies
Term-BL	Term	F2-EXP (i.e., Equation (4))
Concept-BL	Concept (i.e., Section 3.3)	F2-EXP (i.e., Equation (4))
TSConcept-BL	Task specific concept ((Limsopatham et al., 2013d))	F2-EXP (i.e., Equation (4))
Unified	Concept (i.e., Section 4.2)	F2-EXP + Unified (i.e., Equation (5))
Balanced	Concept (i.e., Section 4.3)	F2-EXP + Balanced (i.e., Equation (6))

Table 3: Performance under optimized parameter settings

	Med11		**Med12**	
	MAP	bpref	infNDCG	infAP
Term-BL	0.3474	0.4727	0.4695	0.2106
Concept-BL	0.3967	0.5476	0.5243	0.2497
TSConcept-BL	0.3964	0.5329	0.5283	0.2694
Unified	0.4235^{T}	0.5443^{T}	0.5416^{T}	0.2586^{T}
Balanced	$\mathbf{0.4561}^{T,C,TS}$	$\mathbf{0.5697}^{T,C,TS}$	$\mathbf{0.5767}^{T,C,TS}$	$\mathbf{0.2859}^{T,C,TS}$

The data sets are denoted as *Med11* and *Med12*. Both data sets used the same document collection with 100,866 medical records, each of which is associated with a unique patient visit to the hospital or emergency department. Since the task is to retrieve relevant visits, we merged all the records from a visit to form a single document for the visit, which leads to 17,198 documents in the collection. There are 34 queries in *Med11* and 47 in *Med12*. These queries were developed by domain experts based on the "inclusion criteria" of a clinical study (Voorhees and Tong, 2011; Voorhees and Hersh, 2012).

After applying MetaMap to both documents and queries, we can construct a concept-based collection. Since documents are often much longer, we can first segment them into sentences, get the mapping results for each sentence, and then merge them together to generate the concept-based representation for the documents.

Table 1 compares the statistics of the term-based and the concept-based collections, including the number of unique tokens in the collection (i.e., the number of terms for term-based representation and the number of concepts for concept-based representation), the average number of tokens in the documents (AvgDL) and the average number of tokens in the queries for these two collections (AvgQL11 and AvgQL12). It is interesting to see that the number of unique tokens is much smaller when using the concept-based indexing. This is expected since terms are semantically related and a group of related terms would be mapped to one semantic concept. Moreover, we observe that the document length and query length are similar for both collections. This is caused by the fact that concepts are related and the MetaMap would map an aspect to multiple related concepts.

Table 2 summarizes the methods that we compare in the experiments. Following the evaluation methodology used in the medical record track, we use MAP@1000 as the primary measure for *Med11* and also report bpref. For *Med12*, we take infNDCG@100 as the primary measure and also report infAP@100. Different measures were chosen for these two sets mainly because different pooling strategies were used to create the judgment pools (Voorhees and Hersh, 2012).

5.2 Performance Comparison

Table 3 shows the performance under optimized parameter settings for all the methods over both data sets. The performance is optimized in terms of MAP in *Med11*, and infNDCG in *Med12*, respectively. α and b are tuned from 0 to 1 with the step 0.1. Note that ^{T}, ^{C} and ^{TS} indicate improvement over Term-BL, Concept-BL and TSConcept-BL is statistically significant at 0.05 level based on Wilcoxon signed-rank test, respectively.

Results show that *Balanced* method can significantly improve the retrieval performance over both collections. *Unified* method outperforms the baseline methods in terms of the primary measure on both collections, although it fails to improve the infAP on *Med12* for one baseline method. It is not surprising to see that *Balanced* method is more effective than *Unified* since the former satisfies both of the proposed retrieval constraints while the lat-

Table 4: Testing Performance

Trained on	**Med12**		**Med11**	
Tested on	**Med11**		**Med12**	
Measures	MAP	bpref	infNDCG	infAP
Term-BL	0.3451	0.4682	0.4640	0.2040
Concept-BL	0.3895	0.5394	0.5194	0.2441
TSConcept-BL	0.3901	0.5286	0.5208	0.2662
Unified	$0.4176^{T,C}$	0.5391^{T}	0.5346^{T}	0.2514^{T}
Balanced	$\mathbf{0.4497}^{T,C,TS}$	$\mathbf{0.5627}^{T,C,TS}$	$\mathbf{0.5736}^{T,C,TS}$	$\mathbf{0.2811}^{T,C,TS}$

ter satisfies only one. Finally, we noticed that the performance difference between TSConcept-BL and Concept-BL is not as significant as the ones reported in the previous study (Limsopatham et al., 2013d), which is probably caused by the difference of problem set up (i.e., record-level vs. visit-level as discussed in Section 3.1).

We also conduct experiments to train parameters on one collection and compare the testing performance on the other collection. The results are summarized in Table 4. Clearly, *Balanced* is still the most effective regularization method. The testing performance is very close to the optimal performance, which indicates that the proposed methods are robust with respect to the parameter setting.

Moreover, we would like to point out that the testing performance of *Balanced* is comparable to the top ranked runs from the TREC Medical records track. For example, the performance of the best automatic system in *Med11* (e.g., CengageM11R3) is 0.552 in terms of bpref, while the performance of the best automatic system in *Med12* (e.g., udelSUM) is 0.578 in terms of infNDCG. Note that the top system of *Med12* used multiple external resources such as Wikipedia and Web, while we did not use such resources. Moreover, our performance might be further improved if we apply the result filtering methods used by many TREC participants (Leveling et al., 2012).

Table 5: Selecting representative concepts

	MAP	**bpref**
Unified (i.e., Unified-max)	**0.4235**	**0.5443**
Unified-min	0.3894	0.5202
Unified-avg	0.4164	0.5303

5.3 More Analysis

In the *Unified* method, we chose the concept with the maximum IDF as the representative concept among all the variants. We now conduct experiments on *Med11* to compare its performance with those of using average IDF and minimum IDF ones as the representative concept. The results are shown in Table 5. It is clear that using maximum IDF is the best choice, which is consistent with our observation from the data exploratory analysis shown in Figure 2.

Table 6: Estimating query aspect importance

	MAP	**bpref**
Balanced (i.e., Balanced-max)	**0.4561**	**0.5697**
Balanced-min	0.4216	0.5484
Balanced-avg	0.4397	0.5581

Table 7: Regularization components in *Balanced*

	MAP	**bpref**
Balanced	**0.4561**	**0.5697**
Confidence only	0.4294	0.5507
Importance only	0.4373	0.5598

In the *Balanced* method, we used the maximum IDF value to estimate the query importance. We also conduct experiments to compare its performance with those using the minimum and average IDF values. Table 6 summarizes the results, and shows that using the maximum IDF value performs better than the other choices.

As shown in Equation (3), the *Balanced* method regularizes the weights through two components: (1) normalized confidence score of each aspect, i.e., $\frac{\sum_{e' \in Q} i(e')}{\sum_{e'' \in \mathcal{A}(e)} i(e'')}$; and (2) the importance of the query aspect, i.e., $Imp_A(\mathcal{A}(e))$. To examine the effectiveness of each component, we conduct experiments using the modified *Balanced* method with only one of the components. The results are shown in Table 7. It is clear that both components are essential to improve the retrieval performance.

Finally, we report the performance improvement of the proposed methods over the *Concept-BL* for each query in Figure 3. Clearly, both of the

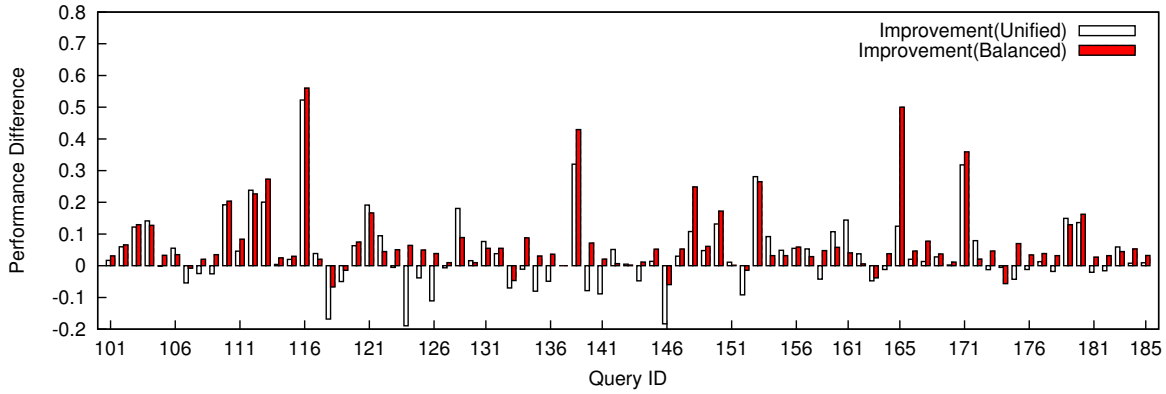

Figure 3: Improvement of proposed methods (Compared with the Concept-BL method).

proposed methods can improve the effectiveness of most queries, and the *Balanced* method is more robust than the *Unified* method.

6 Conclusions and Future Work

Medical record retrieval is an important domain-specific IR problem. Concept-based representation is an effective approach to dealing with ambiguity terminology in medical domain. However, the results of the NLP tools used to generate the concept-based representation are often not perfect. In this paper, we present a general methodology that can use axiomatic approaches as guidance to regularize the concept weighting strategies to address the limitations caused by the inaccurate concept mapping and improve the retrieval performance. In particular, we proposed two weighting regularization methods based on the relations among concepts. Experimental results show that the proposed methods can significantly outperform existing retrieval functions.

There are many interesting directions for our future work. First, we plan to study how to automatically predict whether to use concept-based indexing based on the quality of MetaMap results, and explore whether the proposed methods are applicable for other entity linking methods. Second, we will study how to leverage other information from knowledge bases to further improve the performance. Third, more experiments could be conducted to examine the effectiveness of the proposed methods when using other ranking strategies. Finally, it would be interesting to study how to follow the proposed methodology to study other domain-specific IR problems.

References

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*.

Alan R. Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *JAMIA*, 17(3):229–236.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of AMIA Symposium*.

Practice Management Information Corporation. 1999. *ICD-9-CM: International Classification of Diseases, 9th Revision, Clinical Modification, 5th Edition*. Practice Management Information Corporation.

Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell Loane, Francois Lang, James G. Mork, Nicholas Ide, and Alan R. Aronson. 2012. NLM at TREC 2012 Medical Records Track. In *Proceedings of TREC 2012*.

Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of SIGIR'05*.

Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Proceedings of SIGIR'04*.

Charles P. Friedman, Adam K. Wong, and David Blumenthal. 2010. Achieving a nationwide learning health system. *Science Translational Medicine*.

Beval Koopman, Michael Lawley, and Peter Bruza. 2011. AEHRC & QUT at TREC 2011 Medical Track : a concept-based information retrieval approach. In *Proceedings of TREC'11*.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, Deanne Vickers, Luke Butt, and Peter D. Bruza.

2012. 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. In *Proceedings of TREC'12*.

Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, and Gareth J. F. Jones. 2012. DCU@TRECMed 2012: Using adhoc Baselines for Domain-Specific Retrieval. In *Proceedings of TREC 2012*.

Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013a. Inferring conceptual relationships to improve medical records search. In *Proceedings of OAIR'13*.

Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013b. Learning to combine representations for medical records search. In *Proceedings of SIGIR'13*.

Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013c. Learning to selectively rank patients' medical history. In *Proceedings of CIKM'13*.

Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013d. A task-specific query and document representation for medical records search. In *Proceedings of ECIR'13*.

Jimmy Lin and Dina Demner-Fushman. 2006. The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 99–106, New York, NY, USA. ACM.

Carolyn E Lipscomb. 2000. Medical Subject Headings (MeSH). *The Medical Library Association*.

Christopher D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Bridget T. McInnes, Ted Pedersen, and Serguei V. S. Pakhomov. 2009. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proceedings of AMIA Symposium*.

Yanjun Qi and Pierre-Francois Laquerre. 2012. Retrieving Medical Records: NEC Labs America at TREC 2012 Medical Record Track. In *Proceedings of TREC 2012*.

S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. 1996. Okapi at TREC-3. pages 109–126.

Charles Safran, Meryl Bloomrosen, W. Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C. Tang, and Don E. Detmer. 2007. White paper: Toward a national framework for the secondary use of health data: An american medical informatics association white paper. *JAMIA*, 14(1):1–9.

Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of SIGIR'96*.

Ellen M. Voorhees and William Hersh. 2012. Overview of the TREC 2012 Medical Records Track. In *Proceedings of TREC 2012*.

Ellen M. Voorhees and Richard M. Tong. 2011. Overview of the TREC 2011 Medical Records Track. In *Proceedings of TREC 2011*.

Xin Yan, Raymond Y.K. Lau, Dawei Song, Xue Li, and Jian Ma. 2011. Toward a semantic granularity model for domain-specific information retrieval. *ACM TOIS*.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of SIGIR'01*.

Wei Zheng and Hui Fang. 2010. Query aspect based term weighting regularization in information retrieval. In *Proceedings of ECIR'10*.

Wei Zhou, Clement Yu, Neil Smalheiser, Vetle Torvik, and Jie Hong. 2007. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proceedings of SIGIR'07*.

Dongqing Zhu and Ben Carterette. 2012. Combining multi-level evidence for medical record retrieval. In *Proceedings of SHB'12*.