Improving Health Records Search Using Multiple Query Expansion Collections

Donqqing Zhu and Ben Carterette
Department of Computer and Information Sciences
University of Delaware
Newark, DE 19716, USA
Email: {zhu,carteret}@cis.udel.edu

***Abstract*—The increasing prevalence of electronic health records (EHR), along with the needs for enhanced clinical care, presents new challenges to information retrieval (IR). Many clinical decision-making tasks following the philosophy of Evidence-Based Medicine (EBM) rely on the ability to find relevant health records and gather sufficient clinical evidence under severe time constraints.**

In this work, we present a system built upon statistical IR methods for searching flat-text health records (i.e. the doctors' notes sections of EHR) for patients with particular conditions specified via a keyword query. In particular, we use multiple external repositories for query expansion, and introduce two novel model weighting methods. Cross-validation results show that our system improves a strong baseline by 30% on mean average precision (MAP), and has a promising overall performance when compared with a manual system doing the same task.

***Keywords*-information retrieval; electronic health records; query expansion**

I. Introduction

The increasing prevalence of electronic health records containing a full record of a patient's health and physical condition has the potential to transform research in health and medicine. Doing so will require novel computational methods, particularly from information retrieval (IR), data and text mining, and machine learning.

One specific task for IR technology is *systematic reviews*, a study of a sample of patients that all have a particular condition. Running a systematic review necessarily first involves a search for patients to include. While some of these searches can be straightforward (practically solvable with SQL-like queries on relational data), many run up against the standard problems of information retrieval: heterogeneity of language, polysemy, synonymy, and all the other problems that computational processing of natural language presents. For example, a query for "patients with hearing loss" will match many nonrelevant documents with references to weight loss, simply because the word "loss" occurs much more often with "weight" than with "hearing".

In this work, we present a system built upon statistical IR methods for searching flat-text health records (i.e. the doctors' notes sections of EHR) for patients with particular conditions specified via a keyword query. We leverage information from multiple external repositories, ranging from general web collections to medical-specific thesauri, to improve the baseline system. In particular, we introduce two novel weighting methods for our retrieval models. As our results will show, these two weighting approaches significantly improve the retrieval effectiveness over their un-weighted counterparts. Our final system, incorporating three additional medical features, improves a strong baseline by about 30% on MAP.

II. Related Work

As EHR become more prevalent, attempts have been made to transfer search engine technology to EHR retrieval for various applications [1]. The EMERSE (Electronic Medical Record Search Engine) system, as one of the earliest and successful non-commercial EHR search engines, supports free-text queries, and has been used by medical professionals in a few hospitals, health centers, and clinics since its initial introduction in 2005 [1], [2]. Though EMERSE has not achieved widespread adoption and there is little discussion about its search algorithms, a few interesting research work have been done using EMERSE: Seyfried et al. [2] compared EMERSE-facilitated chart reviews with manual reviews, and concluded that using a well-designed EHR search engine for retrieving information in free-text EHR can provide significant time saving while preserving reliability.

Yang et al. [3] analyzed a query log of EMERSE over the course of 4 years , and found that the coverage of EHR query terms by a meta-dictionary is much lower than the usual 85-90% coverage of Web queries by English dictionaries. Thus, they suggested seeking beyond the use of medical ontologies to enhance medical information retrieval.

Apart from these few attempts, methods emerging from research on IR have not been well explored, largely due to the sensitivity of patient data, preventing its use by academic researchers. Fortunately, the Text REtrieval Conference (TREC) initiated a Medical Records track in 2011 making a set of real medical records and human judgments of relevance to search queries available to the research community.

TREC is an annual evaluation workshop/competition with the goal of providing a common experimental setting for researchers that want to work on particular search tasks. Each year, there are up to 7 "tracks" devoted to a different

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

search task. Organizers provide documents and information needs to participants, ensuring that all participants are using the same data and working towards the same task. TREC organizers also oversee the collection of human relevance judgments, which are instrumental in understanding the effectiveness of a search system.

Most participants in TREC's Medical Records track tried using medical-specific knowledge to enhance retrieval, but only a few of them achieved positive results. King et al. [4], Goodwin et al. [5], and Zhu et al. [6] leveraged information from medical-related external sources, such as UMLS, SNOMED-CT, and PubMed query log. They all obtained large improvement over their baselines which used no medical-specific knowledge. Daoud et al. [7], Schuemie et al. [8], and Wu et al. [9] used similar sources, such as UMLS, MeSH, RxNorm, DrugBank, etc. However, they all obtained very little or no improvement over their baselines.

In fact, all the above groups used medical knowledge in very similar ways, yet the outcomes varied a lot. This is most likely because they used different underlying retrieval models that interact with medical domain knowledge in very different ways. Thus, in this work, we focus on improving retrieval models that can interact well with medical knowledge. In particular, we seek novel weighting methods for combining information from external resources to improve the effectiveness of query expansion.

III. Retrieval Task and Data

In this section, we describe the retrieval task and several data sources as listed in Table I for our experiments.

A. Target Collection and Retrieval Task

The target collection (indicated by * in Table I) is the official test collection for the TREC 2011 Medical Records Track [10]. It contains 100,866 de-identified medical reports from the University of Pittsburgh NLP Repository. The retrieval task[1] is an ad hoc search task for patient visits meaning the system must be able to effectively find relevant patient visits for arbitrary user queries. A patient visit to the hospital usually generates multiple medical reports.

Table I
Collection Statistics.

Collection	# documents	vocabulary size	avg doc length
Medical*	100,866	10^5	423
MeSH	n/a	n/a	n/a
ImageCLEF	5,704	10^5	6,495
Genomics	162,259	10^7	6,595
Wikipedia	5,957,529	10^6	1,305
ClueWeb09	44,262,894	10^7	756

[1]http://www-nlpir.nist.gov/projects/trecmed/2011/tm2011.html

1) Medical Reports: Each medical report is an XML file with a fixed set of fields[2]. We will mainly use information from two diagnosis fields that contain ICD-9 codes, and "report_text" field which contains clinical narratives.

As a pre-processing step, we merge reports from a single visit into a visit document thereby converting 100,866 medical reports into 17,198 visit documents based on the report-to-visit mapping information provided with the TREC test collection. By doing so, we combine evidence from scattered reports for each unique visit. Then, we build an Indri index for the merged documents. Indri[3] is an open-source system for indexing and retrieving full-text documents. It supports basic keyword queries, but also has a complex querying language that offers much greater expressive power.

2) Topics: 35 information needs (or "topics" in TREC terminology) were developed by TREC assessors. These needs were designed to require information from the free-text fields, i.e., topics are not answerable solely by the diagnostic codes. Topics are meant to reflect the types of queries that might be used to identify cohorts for comparative effectiveness research [10]. Table II lists several TREC topics as examples. The topic usually specifies the patient's condition, disease, treatment, etc. Relevance judgments for the topics were also developed by TREC assessors based on the pooled results from TREC participants.

Table II
Example topics of medical records track

ID	Topic
107	Patients with ductal carcinoma in situ (DCIS)
118	Adults who received a coronary stent during an admission
112	Female patients with breast cancer with mastectomies during admission

B. External Collections

In addition to the medical records that are the target of retrieval, we leverage information in several other large, widely-available collections: ImageCLEF 2009 Medical Image Retrieval Task dataset [11], TREC 2007 Genomics Track dataset [12], TREC 2009 ClueWeb09 Category B dataset[4](excluding Wikipedia pages), a Wikipedia dataset (containing those excluded Wikipedia pages), and the 2012 Medical Subject Headings (MeSH). Table I provides detailed information about these datasets.We choose these collections because there are existing topics and relevance judgments for analysis and because we want to compare the effects of different sources on retrieval performance.

[2]http://www.dbmi.pitt.edu/nlp/report-repository

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

[4]ClueWeb09 (http://lemurproject.org/clueweb09.php/) was created to support research on IR. It contains about 1 billion web pages, and is used by several tracks of the TREC conference. TREC Category B contains first 50 million English pages including about 6 million Wikipedia pages.

IV. Retrieval Methods

In this section, we describe our retrieval models, including the baseline document scoring method, methods for query expansion, and a novel method for expansion weighting.

A. Retrieval Models

We start with a basic "bag-of-words" probabilistic model: the query likelihood language model. This model scores documents for queries as a function of the probability that query terms would be sampled (independently) from an urn containing all the words in that document. Formally, the scoring function is a sum of the logarithms of smoothed probabilities:

$$\text{score}(D,Q) = \log P(Q|D) = \sum_{i=1}^{n} \log \frac{tf_{q_i,D} + \mu \frac{tf_{q_i,C}}{|C|}}{|D| + \mu}, \quad (1)$$

where q_i is the ith query term, $|D|$ and $|C|$ are the document and collection lengths in words respectively, $tf_{q_i,D}$ and $tf_{q_i,C}$ are the document and collection term frequencies of q_i respectively, and μ is the Dirichlet smoothing parameter. The Indri retrieval system supports this model by default.

The above model is a strong baseline, but the only information it uses is terms in the query and terms in the document. It can be improved by expanding the query with additional "related" terms. These related terms can be derived from a *relevance model*, which is usually built upon top-ranked documents for the query in the target collection. We will describe how to estimate a relevance model in Section IV-B below.

Relevance modeling can be further improved upon by leveraging information in other document collections. Specifically, following Diaz and Metzler [13], we can form relevance models for two or more additional collections, then expand the query using those models.

To achieve better performance, we linearly interpolate the mixture of relevance models with the maximum likelihood (ML) query estimate by formulating the equation:

$$P(w|\theta_Q) = \lambda_Q \frac{\#(w,Q)}{|Q|} + (1-\lambda_Q) \sum_C \lambda_C P(w|\hat{\theta}_{Q,C}), \quad (2)$$

where the first part is the weighted ML query estimate and the second part represents the mixture of relevance models. In particular, $P(w|\hat{\theta}_{Q,C})$ is the estimate of relevance model based on expansion collection C. λ's are collection weights and $\sum_C \lambda_C = 1$. Indri naturally supports such queries with the "#weight" operator; we implement Eq.2 in Indri by formulating a query of the following format:

#weight(
  λ_Q #combine(w_1 w_2 ... $w_{|Q|}$)
  $(1-\lambda_Q)$ #weight(
    λ_{C_1}#weight(p_{11} e_{11} p_{12} e_{12} ... p_{1m} e_{1m})
    ...
    λ_{C_n}#weight(p_{n1} e_{n1} p_{n2} e_{n2} ... p_{nm} e_{nm}))).

Here w_i represents a term in the original user query; e_{ij} represents the jth expansion term (in decreasing order of probability p_{ij}) from collection i, n is the number of expansion collections, and m is the number of terms to expand with. The "#weight(p_{i1} e_{i1} p_{i2} e_{i2} ... p_{im} e_{im})" phrase corresponds to the estimate of relevance model $P(w|\hat{\theta}_{Q,C_i})$.

When using a single expansion collection (i.e., $n = 1$ and $\lambda_{C_1} = 1$), Eq.2 reduces to:

$$P(w|\theta_Q) = \lambda_Q \frac{\#(w,Q)}{|Q|} + (1-\lambda_Q) P(w|\hat{\theta}_{Q,C}), \quad (3)$$

We will explain how to obtain expansion terms e and estimate their weights p in the following section.

B. Query Expansion

To implement query expansion we need a way to estimate term probabilities p_{ij} and expansion weights λ_{C_i}. In this section we describe two approaches, one for collections of full-text documents, the other for medical ontologies.

1) General Expansion: For collections containing full-text articles (i.e., all collections in Table I except MeSH), we obtain expansion terms e based on the formula:

$$p_i = P(e_i|\hat{\theta}_{Q,C}) = \sum_{j=1}^{k} \exp\{\frac{tf_{e_i,D_j}}{|D_j|} + \log \frac{|C|}{df_{e_i,C}} + \text{score}(D_j,Q)\}, \quad (4)$$

where $\text{score}(D_j,Q)$ is the query likelihood score for the top jth feedback document in the initial retrieval set ranked by Eq.1, tf_{e_i,D_j} is the term frequency of e_i in document D_j, $df_{e_i,C}$ is the document frequency of e_i in collection C, and $|D_j|$ and $|C|$ are document and collection lengths in words respectively. This formula estimates the importance of term e_i based on its term frequency, inverse document frequency, and scores of the top k documents ranked for the query Q. The m terms with highest probability p are selected as expansion terms.

2) Medical Thesaurus-based Expansion: Medical thesaurus-based expansion differs from general expansion in that there are no feedback documents for obtaining expansion terms e and estimating weights p. Thus, we extract medical concepts from the query for expansion, and propose a novel concept weighting method based on information from a query log. In this work, we use MeSH for query expansion, and call this method *MeSH expansion*, which can be summarized in four steps:

1) Concept identification: use PubMed e-utility to identify MeSH concepts in the query
2) Concept expansion: expand a detected MeSH concept by its entry terms and decedent nodes down level l in the MeSH trees, i.e., obtaining terms e
3) Concept weighting: for each MeSH concept, estimate weight p for e using a PubMed query log

4) Concepts aggregation: merge lists of expansion terms for each concept into one final expansion list

In Step 2, we model term proximity by MeSH concepts. For instance, for MeSH terms "Usher Syndromes" and "Hearing Loss, High-Frequency", we will formulate "#1(usher syndromes)" and "#uw16(#1(hearing loss), high-frequency)" respectively in Indri as expansion terms. The former means "usher syndromes" must occur as a phrase while the latter means "high-frequency" and "hearing loss" can occur within a text window of 16 words. Note that we avoid expanding MeSH concepts by their ancestor nodes because broader concepts are more likely to compromise precision and cause query drift. Moreover, we do not split phrase concepts into single terms because single terms are likely to be semantically different or far less discriminative than their associated phrase concepts (e.g., "usher syndromes", "back pain", "sleep walking disorder", etc.).

The PubMed query log used in Step 3 contains 2,996,301 queries submitted by 627,455 different users [14]. We estimate weight of term e_i by:

$$p_i = \frac{\log N_{e_i,G}}{\sum_j \log N_{e_j,G}}, \tag{5}$$

where $N_{e_i,G}$ is the number of users whose queries contain e_i in query log G. The logarithm dampens the effect of large differences in counts. Eq.5 estimates the popularity of e_i and its variants among users who use them interchangeably to express a medical concept in general. For instance, "hearing impairment" is more common than "hypoacusis" for expressing the concept "hearing loss" and consequently gets a larger weight.

C. A Novel Collection Weighting Method

Applying Eq.4 and Eq.5 to the collections in Table I gives up to six sets of expansion terms (or #weight clauses in the Indri query). Some of these are likely to be better for query expansion than others. Furthermore, while expansion collections with different contents can offer complimentary expansion terms, they can also bring more noise. More importantly, the performance of different expansion collections can vary a great deal across queries, which presents the challenge of assigning appropriate *collection weights* (i.e., λ_C in Eq.2 which estimates how good collection C is in terms of providing good expansion terms).

Here we propose a query-adaptive collection weighting strategy. It is based on the hypothesis that a good expansion collection with respect to a specific query will provide an expansion query E (i.e., a set of weighted expansion terms) that, if used alone for searching against the target collection, will retrieve a set of documents S_E whose contents are similar to the set of documents S_Q obtained by the original query Q. Thus, by measuring the similarity between θ_E and θ_Q (the smoothed uni-gram language models built on S_E and S_Q respectively based on Eq. 1), we can estimate the effectiveness of E for expansion with respect to Q.

In this work, we use the Jensen-Shanon divergence (JSD) for measuring the similarity between θ_E and θ_Q as follows:

$$J(\theta_E||\theta_Q) = \frac{1}{2}(K(\theta_E||\theta_M) + K(\theta_Q||\theta_M)), \tag{6}$$

where $\theta_M(w) = 0.5 \times (\theta_p(w) + \theta_q(w))$ for each word w in the collection, and K is the Kullback-Leibler divergence (KLD): $K(\theta_E||\theta_M) = \sum_w \theta_E \log \frac{\theta_E}{\theta_M}$. JSD is a symmetric version of KLD and is preferred over other distance measures such cosine distance [15]. Smaller JSD score means higher similarity. Thus we use the normalized inverse of JSD score as the expansion weight:

$$\lambda_{C_i} = \frac{J_i^{-1}}{\sum_j J_j^{-1}}, \tag{7}$$

where J_i is JSD score for the ith expansion collection.

Note that in the medical thesaurus-based expansion we propose a term weighting approach while here for using multiple expansion collections we are proposing a collection weighting approach.

V. Evaluation

In this section, we describe evaluation metrics and experimental setup.

A. Evaluation Metrics

We use P10 and bpref (two official evaluation metrics for the 2011 TREC Medical Records track), and MAP (one of the most standard evaluation measures among TREC community) as our evaluation metrics:

1) P10 (precision at rank 10) measures the proportion of relevant documents among the top 10 retrieved.

2) MAP (mean average precision) provides a single-figure measure of quality across recall levels [16]. If $\{d_1,...,d_j\}$ is the set of relevant documents for an information need $q \in Q$, then MAP is defined as:

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{d \in \{d_1,...,d_j\}} \text{Precision}(\text{rank}(d))}{|\{d_1, ..., d_j\}|}, \tag{8}$$

where Precision(k) is the proportion of relevant documents among the top k retrieved.

3) bpref is defined as:

$$\text{bpref} = \frac{1}{R} \sum_r (1 - \frac{|n \text{ ranked higher than } r|}{min(R, N)}), \tag{9}$$

where R is the number of judged relevant documents, N is the number of judged irrelevant documents, r is a relevant retrieved document, and n is a member of the first R irrelevant retrieved documents. bpref computes a preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant documents. It is based on the relative ranks of judged documents only.

Note that in rest of the paper, when we mention bpref or P10, we are referring to the average score of bpref or P10 over all topics in a run. MAP and bpref will be the primary evaluation measures in this work.

B. Experimental Setup

We use the Porter stemmer and a simple standard medical stoplist [17] for stemming and stopping words in documents and queries during indexing and retrieving. Then we conduct 5-fold cross-validation and use the top 1000 retrieved visits[5] for each query to evaluate our system under different settings. In each iteration, we train our system on 28 queries to obtain the best parameter setting for MAP by sweeping over the parameter space according to Table III, and then generate a ranking for each of the remaining 7 queries based on the trained system. When complete, we have full rankings for all 35 topics as a test set. We evaluate the system based on the MAP, bpref, and P10 over all 35 topics.

The baseline system using Eq.1 has only one free variable μ to train. We fix μ to 10000 for other systems to reduce the training time. For systems using single expansion collections, we train them to obtain λ_Q, k, and m, except for MeSH expansion which only needs to train λ_Q because, unlike general expansion, low-ranked MeSH expansion candidate terms can still be highly related to the original query terms. For systems using multiple expansion collections, we fix k to 50, m to 10 (but 30 for MeSH expansion to cover different concepts in the query) for efficiency, and thus λ_Q will be the only free variable for training.

Table III
PARAMETER SPACE FOR TRAINING.

Parameter	From	To	Step Size
Dirichlet smoothing parameter μ	2000	30000	2000
Model weight λ_Q	0.5	1.0	0.1
# of feedback documents k	10	50	10
# of expansion terms m	10	50	10

We train our systems on MAP though bpref is the primary evaluation metric for 2011 medical records track. This is mainly because training on MAP is most commonly used in IR to improve retrieval performance, and we find that training for high MAP improves the performance of other metrics, while training on bpref typically does not.

To access the statistical significance of differences in the performance of two systems, we perform one-tailed paired t-test for difference in MAP.

VI. RESULTS AND ANALYSIS

In this section, we show and discuss the results of various query expansion methods.

[5]Medical Records track guideline requires each retrieval set contain no more than 1000 visits.

A. MeSH Concept Expansion

We compare the effectiveness of different MeSH expansion settings as listed in Table IV-(a): 1) Entry: using entry terms only and without term weighting (i.e., no Step 3 described in Section IV-B2), 2) Tree1: using tree terms only, tree expansion level $l = 1$, and no weighting, 3) EntryTree1: using both entry and tree terms with $l = 1$, no weighting, and 4) WEntryTree-1: weighted EntryTree1 using PubMed query log, 5) WEntryTree[2-6]: similar to WEntryTree-1 but using different values (i.e., $2 \sim 6$) for l.

In Table IV-(a), we can see that our expansion term weighting method brings significant improvement over all other unweighted versions as well as the baseline: we see nearly 12% improvement over the baseline, and 5-7% over the unweighted version. Increasing expansion level l only slightly improves the retrieval effectiveness.

Table IV
EVALUATION OF QUERY EXPANSION. "$X > S$" MEANS THE MAP DIFFERENCE BETWEEN SYSTEM X AND ANY SYSTEM SPECIFIED IN SET S IS STATISTICALLY SIGNIFICANT. THE STATISTICAL SIGNIFICANCE IS DETERMINED USING ONE-TAILED PARIED T-TEST ON QUERIES AND P-VALUE < 0.05. MESH (S) IS EQUIVALENT TO WENTRYTREE1. NLMMANUAL CONTRIBUTED ITS RUN TO THE JUDGMENT POOLS WHILE OTHERS DID NOT.

System	MAP	Significance	bpref	P10
(a) MeSH Expan.				
Baseline (*B*)	0.353		0.469	0.506
Tree1 (*T1*)	0.368 (+4.2%)		0.484	0.509
Entry (*E*)	0.370 (+4.8%)		0.481	0.553
EntryTree1 (*ET1*)	0.377 (+6.8%)		0.490	0.553
WEntryTree1	0.391 (+10.8%)	>{*B, E, T1, ET1*}	0.496	0.547
WEntryTree2	0.394 (+11.6%)	>{*B, T1, E, ET1*}	0.498	0.556
WEntryTree3	0.395 (+11.9%)	>{*B, T1, E, ET1*}	0.498	0.568
WEntryTree4	0.392 (+11.0%)	>{*B, T1, E, ET1*}	0.497	0.556
WEntryTree5	0.391 (+10.8%)	>{*B, T1, E, ET1*}	0.497	0.556
WEntryTree6	0.391 (+10.8%)	>{*B, T1, E, ET1*}	0.497	0.556
(b) Single Expan.				
Baseline (*B*)	0.353		0.469	0.506
ImageCLEF (*I*)	0.371 (+5.1%)		0.492	0.544
Wikipedia (*W*)	0.376 (+6.5%)		0.500	0.550
ClueWeb09 (*C*)	0.390 (+11%)	>{*B*}	0.513	0.556
MeSH (*S*)	0.391 (+11%)	>{*B, I*}	0.496	0.547
Medical (*M*)	0.393 (+11%)	>{*B*}	0.520	0.535
Genomics (*G*)	0.395 (+12%)	>{*B, W*}	0.524	0.553
(c) Multiple Expan.				
Multiple (*MP*)	0.420 (+19.0%)	>{*I,W,C,S,M,G*}	0.541	0.578
WMult (*WMP*)	0.436 (+23.5%)	>{*MP,I,W,C,S,M,G*}	0.558	0.594
(d) Further Improv.				
MedSearch	0.457 (+30%)	>{*WMP*}	0.583	0.612
NLMManual	0.507		0.658	0.727

B. Single Expansion

Before testing multiple collection weighting, we expand queries with each collection on its own based on Eq.3 to compare individual expansion effectiveness. As we can see in Table IV-(b), ImageCLEF and Wikipedia have comparable improvement over the baseline, though the former is more medical-related, much smaller, and less noisy than the latter. The same situation applies to Genomics and

ClueWeb09. However, Genomics and ClueWeb09 are much larger than ImageCLEF and Wikipedia respectively, and Genomics and ClueWeb09 both have significant improvement over the baseline. Genomics is also significantly better than Wikipedia. Thus, we can infer that expansion effectiveness depends on both the quality (i.e., content similarity to the target collection) and size of the expansion collection.

MeSH expansion is different from general expansion in that it relies on a controlled vocabulary from which expansion terms derived are not as diversified as those from a general expansion collection. For instance, for the query "hearing loss", it is difficult for MeSH to provide related expansion terms such as "cochlear", "noise", "auditory", and "binaural" (top-ranked terms from Genomics), "cerumen", "canals", and "tympanic" (from Medical), "vestibular", "ear", and "stape" (from ImageCLEF). Some of these terms do appear in the MeSH trees at upper levels, however, it is hard to find a link to them, i.e., discriminating them from other unrelated tree nodes. Simply including all visited concepts along the path is likely to cause query drift. Moreover, these terms normally appear in phrase concepts having different meanings than individual terms.

MeSH expansion is quite restrictive, yet is comparable to top performing single expansions and is significantly better than the baseline and ImageCLEF. This is most likely because our MeSH expansion emphasizes modeling term proximity which is a big advantage of any medical thesaurus-based expansion over general expansion. Another merit of MeSH expansion is that, if used properly, it rarely includes bad expansion terms, while we have no control of the quality of each expansion term from general expansions.

C. Multiple Expansion

Finally, we use all collections listed in Table I for query expansion based on Eq.2. We compare different collection weighting methods in Table IV-(c). System MP uses uniform collection weights (i.e., same λ_{C_i}'s) while system WMP uses Eq.7 for estimating weights. As we can see, both MP and WMP significantly improve MAP and bpref scores over any single expansion, indicating that using multiple expansion collections is always better than using just a single collection. Moreover, the MAP difference between MP and WMP is statistically significant, which means that our adaptive collection weighting method works well.

D. Further Improvement

We further improve the system by incorporating additional features as described below:

1) ICD Code Expansion: The ICD codes, though mainly used for billing purposes, may give a high level summary of content in the medical records, and their associated descriptions can provide terms that are likely to be helpful to our retrieval system. Thus, we expand ICD-9 codes in medical reports with their corresponding descriptions.

2) Negation Removal: One distinct feature of narrative clinical reports is that negation phrases are frequently used to claim the absence of certain conditions or symptoms [18], such as "cannot tell", "not clear", "without evidence", etc. Negations may cause retrieval false positives. For instance, a simple IR system will consider a document with the sentence "The patient comes in with episodes of orthopnea and has ruled out for an acute coronary syndrome." as relevant to the query "acute coronary syndrome". Thus, we use ConText[6] [19] to remove all negated portions of the sentences from the medical records before indexing.

3) Age and Gender Filtering: Topics of the medical records track may also have age and gender restrictions to further specify the kind of patients required for the clinical study [10]. Thus, we use simple regular expressions to search in both topics and records for words/phrases that indicate the patient's age (e.g., "children", "adults", and "old patients") and gender (e.g., "women", "his", and "female"). Then, we use this information to further filter from the retrieval set visits that do not meet the age/gender restrictions.

We call this further enhanced system MedSearch as shown in Table IV-(d). MedSearch improves the baseline by about 30% on MAP. For comparison, we also show the evaluation scores of the best manual run from system NLMManual [20] in 2011 Medical Records track. In this run, medical experts are involved in the process of converting a query into a ranked list by repeatedly modifying the query by hand. As we can see that with a big improvement (30% on MAP) over the baseline the performance of MedSearch is quite close to NLMManual. Note that because our system did not contribute runs to TREC judgment pools, our scores could potentially be higher. For instance, P10 score 0.612 is just the lower bound for our system, because there are quite a few visits with missing judgments in our top 10 retrieved (while NLMManual had all the top 10 results judged [10]).

VII. Conclusion and Future Work

In this work, we build a health record search system and improve its retrieval performance by using multiple external collections. In particular, we propose two novel weighting methods to guide query expansion, i.e. medical concept weighting (Section VI-A) and expansion collection weighting (Section VI-C). They both significantly improve retrieval effectiveness comparing with uniform weighting methods. We further enhance our system by incorporating three additional medical features. The final system improves a strong baseline by about 30% on MAP, and presents a promising overall performance when compared with a manual system doing the same task.

For future work, we will incorporate more medical knowledge for query expansion. Our proposed concept weighting method can directly be applied it to other medical thesauri

[6]A negation detection tool available at http://code.google.com/p/negex/.

(e.g., SNOMED-CT, UMLS, etc.). Thus, we will explore using other medical thesauri for query expansion.

REFERENCES

[1] D. A. Hanauer, "EMERSE: The electronic medical record search engine." *AMIA Annual Symposium Proceedings*, vol. 331, no. 7531, p. 941, Jan. 2006.

[2] L. Seyfried, D. Hanauer, and D. Nease, "Enhanced identification of eligibility for depression research using an electronic medical record search engine," *International Journal of Medical Informatics*, vol. 78, no. 12, pp. e13–e18, Dec. 2009.

[3] L. Yang, Q. Mei, K. Zheng, and D. Hanauer, "Query log analysis of an electronic health record search engine," in *AMIA Annual Symposium Proceedings*, 2011, pp. 915–924.

[4] B. King, L. Wang, I. Provalov, and J. Zhou, "Cengage Learning at TREC 2011 medical track," in *Proceedings of The 20th Text REtrieval Conference (TREC)*, 2011.

[5] T. Goodwin, B. Rink, K. Roberts, S. M. Harabagiu, and R. Tx, "Cohort shepherd: Discovering cohort traits from hospital visits," in *Proceedings of The 20th Text REtrieval Conference (TREC)*, 2011.

[6] D. Zhu and B. Carterette, "Using multiple external collections for query expansion," in *Proceedings of The 20th Text REtrieval Conference*, 2011.

[7] M. Daoud, D. Kasperowicz, J. Miao, and J. Huang, "York University at TREC 2011: Medical Records Track," in *Proceedings of The 20th Text REtrieval Conference*, 2011.

[8] M. Schuemie, "DutchHatTrick: Semantic query modeling, ConText, section detection, and match score maximization," in *Proceedings of The 20th Text REtrieval Conference*, 2011.

[9] H. Wu and H. Fang, "An exploration of new ranking strategies for medical record tracks," in *Proceedings of The 20th Text REtrieval Conference*, 2011.

[10] E. M. Voorhees and R. M. Tong, "Overview of the TREC 2011 medical records track," in *Proceedings of The 20th Text REtrieval Conference (TREC)*, 2011.

[11] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C. E. Kahn, and W. R. Hersh, "Overview of the CLEF 2009 medical image retrieval track," in *CLEF (2)*, 2009, pp. 72–84.

[12] W. R. Hersh, A. M. Cohen, L. Ruslen, and P. M. Roberts, "TREC 2007 genomics track overview," in *TREC*, 2007.

[13] F. Diaz and D. Metzler, "Improving the estimation of relevance models using large external corpora," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2006, pp. 154–161.

[14] J. R. Herskovic, L. Y. Tanaka, W. R. Hersh, and E. V. Bernstam, "A day in the life of PubMed: Analysis of a typical day's query log." *JAMIA*, vol. 14, no. 2, pp. 212–220, 2007.

[15] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, "Learning to estimate query difficulty," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 512, 2005.

[16] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, 1st ed. Addison Wesley, Feb. 2009.

[17] W. Hersh, *Information Retrieval: A Health and Biomedical Perspective*, 3rd ed., ser. Health Informatics. Springer, 2009.

[18] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "Evaluation of negation phrases in narrative clinical reports." *Proceedings of AMIA Symposium*, pp. 105–109, Jan. 2001.

[19] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, "Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports," *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 839–851, 2009.

[20] D. Demner-Fushman, S. Abhyankar, A. Jimeno-Yepes, R. Loane, B. Rance, F. Lang, N. Ide, E. Apostolova, and A. R. Aronson, "A knowledge-based approach to medical records retrieval," in *Proceedings of The 20th Text REtrieval Conference*, 2011.