Knowledge-Based Query Expansion to Support Scenario-Specific Retrieval of Medical Free Text

Zhenyu Liu, Wesley W. Chu
UCLA Computer Science Department
Los Angeles, CA 90024
{vicliu,wwc}@cs.ucla.edu

ABSTRACT

In retrieving medical free text, users are often interested in answers relevant to certain scenarios, scenarios that correspond to common tasks in medical practice, e.g., "treatment" or "diagnosis" of a disease. Consequently, the queries they pose are often scenario-specific, e.g., "lung cancer, treatment." A fundamental challenge in handling such queries is that scenario terms in the query (e.g. "treatment") are too general to match specialized terms in relevant documents (e.g. "lung excision"). In this paper we propose a knowledge-based query expansion method that exploits the UMLS knowledge source to append the original query with additional terms that are specifically relevant to the query's scenario(s). We compare the proposed method with statistical expansion that only explores statistical term correlation and expands terms that are not necessarily scenario specific. Our study on the OHSUMED testbed shows that the knowledge-based method which results in scenario-specific expansion is able to improve more than 5% over the statistical method on average, and about 10% for queries that mention certain scenarios, such as "treatment of a disease" and "differential diagnosis of a symptom/disease."

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Knowledge-Based Approach, Medical Free-Text Retrieval, Automatic Query Expansion

1. INTRODUCTION

Recent years have witnessed a phenomenal growth of Web-based medical document collections. Such collections, e.g., PubMed[1] and Harrison's Online,[2] provide comprehensive coverage of medical literature and teaching materials. In searching these collections, it is often desirable to retrieve only those documents pertaining to a specific medical "scenario," where a scenario is typically defined as a frequently-reappearing medical task. For example, in diagnosing a potential lung cancer patient, a physician may pose a query "lung cancer, diagnosis" in order to find out the latest diagnostic techniques about this disease. Here "diagnosis" is the medical task that marks the scenario of this query. Recent studies [1, 2, 3, 4, 5] reveal that in clinical practice, as many as 60% of physicians' queries center around a limited number of scenarios, e.g. "treatment," "diagnosis," "etiology," etc. While the contextual information in such queries (e.g., the particular disease of a patient such as "lung cancer," the age group of that patient, etc.) varies from case to case, the set of frequently-asked medical scenarios remains unchanged. Retrieving documents that are specifically related to the query's scenario is referred to as *scenario-specific retrieval*.

Scenario-specific retrieval is not adequately addressed by traditional text retrieval systems (e.g. SMART [6] or INQUIRY [7]). Such systems suffer from the fundamental problem of *query-document mismatch* [8] when handling scenario-specific queries. Scenario terms in these queries are typically general, e.g., "treatment" in the query "lung cancer, treatment," while full-text medical documents often discuss the same topic using much more specialized terms, e.g., "lung excision" or "chemotherapy." Such general scenario terms fail to match with the specialized terms in relevant documents, resulting in poor retrieval performance. Because of such ineffectiveness, searching online document collections for clinical usage is still frustrating, labor-intensive and time-consuming, as reported by recent studies [9, 10, 11, 2, 3]. Although about one third of a physician's clinical questions can potentially be answered by such online information resources [12], the overall usage of them in medical practice remains relatively low [1, 3].

There has been recent research on *query expansion* [13, 14, 15, 16] to ameliorate the query-document mismatch problem. However, such techniques also have difficulties handling scenario-specific queries. In principle, query expansion techniques append the original query with specialized terms that have a statistical co-occurrence relationship with original query terms in medical literature. Although appending such specialized terms makes the expanded query a better match with relevant documents, the expansion is not scenario-specific. For example, in handling the query "lung cancer, treatment," existing query expansion techniques will append not only terms such as "lung excision" or "chemotherapy" that are relevant to the "treatment" scenario, but also irrelevant terms like "smoking" and "lymph node," simply because the latter terms co-occur with "lung cancer" in medical literature. Appending non-scenario-specific terms leads to the retrieval of documents that are irrelevant to the original query's scenario, diverging from our goal of scenario-specific retrieval.

[1] `http://www.ncbi.nlm.gov/entrez/query.fcgi?db=PubMed`
[2] `http://harrisons.accessmedicine.com`

SAC'05 March 13-17, 2005, Santa Fe, New Mexico, USA

In this paper, we propose a *knowledge-based query expansion* technique to support scenario-specific retrieval. Our technique exploits domain knowledge in order to restrict query expansion to scenario-specific expansion terms, thus improving upon traditional query expansion approaches. The following are challenges in developing such a knowledge-based technique:

- **Using domain knowledge to automatically identify scenario-specific expansion terms.** It is impractical to ask users or domain experts to manually identify scenario-specific terms for every query and all possible scenarios, and therefore an automatic approach is highly desirable. However, the distinction between scenario-specific expansion terms and non-scenario-specific ones may seem apparent to a human expert, but can be very difficult for a program. To address this problem, we propose a method that exploits a domain-specific knowledge source to treat this distinction.
- **Incompleteness of knowledge sources.** Knowledge sources are usually not specifically designed for the purpose of scenario-specific retrieval. As a result, scenarios frequently appearing in medical queries may not be adequately supported by those knowledge sources. To address this problem, we propose a knowledge-acquisition methodology to supplement the existing knowledge sources with additional knowledge that supports undefined scenarios.

The rest of this paper is structured as follows. A framework for knowledge-based query expansion is presented in Section 2, and detailed methods in this framework are described in Section 3. We experimentally evaluate the framework and report the results in Section 4. Section 5 discusses related works and Section 6 concludes the paper.

2. A FRAMEWORK FOR KNOWLEDGE-BASED QUERY EXPANSION

Figure 1 depicts the components in a knowledge-based query expansion and retrieval framework. Given an original query, *Statistical Query Expansion* (whose scope is marked by the inner dotted rectangle) will first derive *candidate expansion concepts*[3] that are statistically co-occurring with the original query concepts (Section 3.1), and assign weights to each candidate concept according to the statistical co-occurrence. Such weights will be carried through the framework.

Based on the candidate concepts derived by statistical expansion, *Knowledge-based Query Expansion* (whose scope is marked by the outer dotted rectangle) further derives the scenario-specific expansion concepts, with the aid of domain knowledge such as UMLS [19] (Section 3.2). Such knowledge may be incomplete to include all possible query scenarios. Therefore, in an off-line process, we use a *Knowledge Acquisition and Supplementation* module to supplement the incomplete knowledge (Section 3.3).

After the query is expanded with scenario-specific concepts, we use a *Vector Space Model* (VSM) to compare the similarity between the expanded query and each document, and further output the top-ranked documents.

3. METHOD

In this section, we first describe existing methods to derive statistically-related concepts. Afterwards, we propose a knowledge-based method to automatically detect scenario-specific concepts among those statistically-related concepts. This addresses the first challenge identified in the introduction section. In the end, we describe a knowledge-acquisition methodology to supplement the incomplete knowledge source so as to handle previously unsupported scenarios, which addresses the second challenge.

3.1 Deriving statistically-related expansion concepts

Statistical expansion is also referred to as *automatic query expansion* [8, 16]. The basic idea is to derive concepts that are statistically related to the original query concepts in a document collection (e.g. OHSUMED [20]). Appending such concepts to the original query makes the query expression more specialized and helps the query better match with relevant documents. Depending on how such statistically-related concepts are derived, statistical expansion methods fall into two major categories:

- *Co-occurrence-thesaurus-based expansion* [13, 14, 15]. In this method, a *concept co-occurrence thesaurus* is first constructed automatically offline. Given a vocabulary of M concepts, the thesaurus is an $M \times M$ matrix, where the $\langle i, j \rangle$ element quantifies the co-occurrence between concept i and concept j. When a query is posed, we look up the thesaurus to find all concepts that statistically co-occur with concepts in the given query, and assign weights to those co-occurring concepts according to the values in the co-occurrence thesaurus. A detailed procedure to compute the co-occurrence thesaurus and to assign weights to expansion concepts can be found in [13].
- *Pseudo-relevance-feedback-based expansion* [21, 22, 23, 24, 16]. In pseudo relevance feedback, the original query is used to perform an initial retrieval. Concepts extracted from top-ranked documents in the initial retrieval are considered statistically related and are appended to the original query. This approach resembles the well-known *relevance feedback* approach [25, 26] except that, instead of asking users to identify relevant documents as feedbacks, top-ranked (e.g. top-10) documents are automatically treated as "pseudo" relevant documents, and subsequently inserted into the feedback loop. Weight assignment in pseudo relevance feedback [22] typically follows the same weighting scheme ($\langle \alpha, \beta, \gamma \rangle$) as conventional relevance feedback techniques [25].

We note that the choice of statistical expansion method is orthogonal to the design of the knowledge-based expansion framework (Figure 1). In our current experimental evaluation, we use the co-occurrence-thesaurus-based method as described in [13] to derive statistically-related concepts.

3.2 Deriving scenario-specific expansion concepts

Using the method in the previous section we derive candidate expansion concepts that are statistically related to the original query. Only a sub-set of these candidate concepts is relevant to the original query's scenario. In this section we first present a knowledge-based method to select such scenario-specific concepts. Further we discuss how to adjust the weights of these selected scenario-specific concepts to increase their significance in the expanded query.

A knowledge-based method to identify scenario-specific expansion concepts. The basic idea of our knowledge-based method is the following: A scenario-specific query consists of two parts: a key concept c_k (e.g., "lung cancer") and several scenario concepts c_s's (e.g., "treatment," "diagnosis," etc.). Given a scenario-specific query in free-text format, we can detect c_k using concept indexing methods existing in the literature, e.g., IndexFinder [18],

[3] In the rest of this paper, a concept is referred to as a word or a word phrase that has a concrete meaning in a particular application domain. In the medical domain, concepts in free text can be extracted using existing tools, e.g. MetaMap [17], IndexFinder [18], etc.

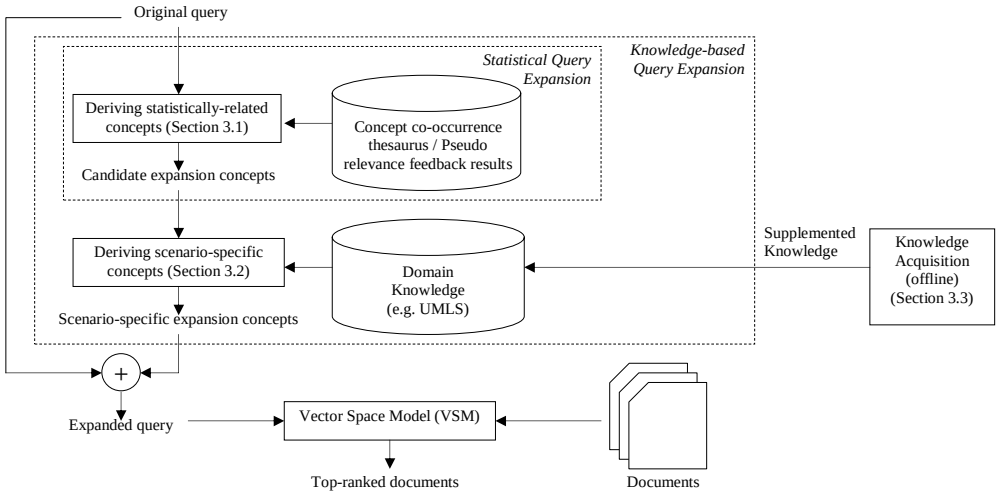


Figure 1: A knowledge-based query expansion and retrieval framework

MetaMap [17], etc. The scenario concepts can be indicated by the user by selecting from a list of scenarios, since the number of frequently-asked scenarios is limited.

Using statistical expansion, we obtain candidate expansion concepts co-occurring with the key concept c_k, e.g., "smoking," "lung excision," etc., for c_k ="lung cancer." Afterwards, we explore a domain-specific knowledge source to identify possible relationships between each candidate expansion concept and c_k. For example, the knowledge source may indicate that "smoking" is a "risk factor" for "lung cancer," whereas "lung excision" is a "treatment" method for this disease. Among these identified relationships, certain relationships are "desirable" because they match with scenarios of the original query. Thus, our knowledge-based method will keep only the candidate concepts that have a desirable relationship with c_k. Since such concepts should be specifically relevant to the original query's scenarios, appending such concepts should lead to scenario-specific expansion.

To develop the idea above in full details, in the following, we first introduce the knowledge structure used in our study, and then describe our knowledge-based method as a 5-step procedure.

For free text retrieval in the medical domain, we choose UMLS to be our domain-specific knowledge. UMLS is a comprehensive medical knowledge source developed by the National Library of Medicine (NLM) [19]. It consists of the following major components: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon, and our method relies on the first two components. The Metathesaurus contains over 800,000 medical concepts (small circles in Figure 2). A group of concepts (enclosed by a dotted circle in Figure 2) in the Metathesaurus belong to a Semantic Type (rectangles in Figure 2) in the Semantic Network. For example, "lung cancer" and other disease concepts belong to one Semantic Type called "Disease or Syndrome." The Semantic Network is modelled as an Entity-Relation diagram in which each Semantic Type is an entity and Semantic Types are associated via relationships. In Figure 2, for example, Semantic Types "Therapeutic and Preventive Procedures," "Medical Device" and "Pharmacologic Substance" have a "treats" relationship with Semantic Type "Disease or Syndrome."

Given this knowledge structure, we propose the following procedure to identify the scenario-specific expansion concepts:

1. We identify the key concept c_k in the scenario-specific query and locate its position in the Metathesaurus.
2. We navigate from c_k to the Semantic Type it belongs to (e.g. from "lung cancer" to "Disease or Syndrome" in Figure 2).
3. Starting from c_k's Semantic Type, we follow the relationships as indicated by the scenario concepts c_s's, e.g., following "treats" if a c_s is "treatment," and reach a set of relevant Semantic Types (e.g., "Medical Device," "Therapeutic or Preventive Procedure" and "Pharmacologic Substance" in Figure 2).
4. Among all these candidate expansion concepts derived by statistical expansion (Section 3.1), those concepts that belong to these relevant Semantic Types are selected as scenario-specific expansion concepts (e.g., the shaded circular areas in Figure 2).
5. Concepts in the Metathesaurus are interconnected by parent-child relationships, forming a general-to-specific concept hierarchy (which is not displayed in Figure 2). To match relevant documents discussing concepts that are more general or specialized than the key concept c_k, we add c_k's surrounding concepts in this hierarchy (parents, children and sibling con-

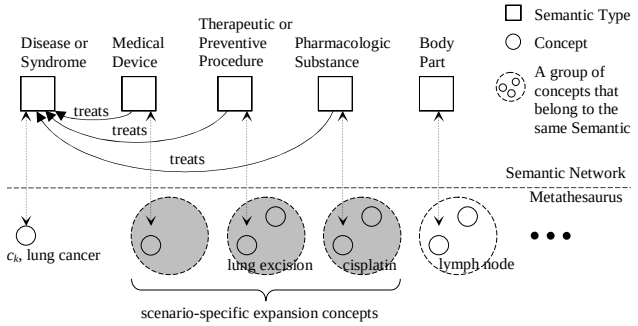


Figure 2: A knowledge-based method to identify scenario-specific expansion concepts

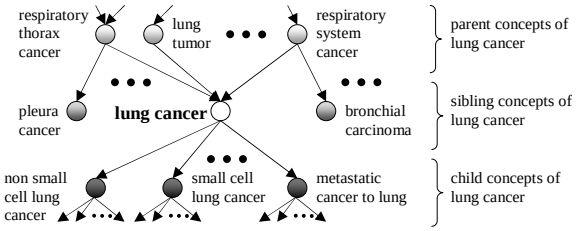


Figure 3: The parent, child and sibling concepts surrounding "lung cancer" as defined by the UMLS Metathesaurus

Concept	Weight
nonsmall cell lung cancer	2.77
large cell carcinoma	2.28
small cell lung cancer	2.15
cancer	1.84
radon daughters	1.76
mediastinal lymph node	1.71
non small cell	1.68
lung carcinoma	1.64
sputum cytology	1.62
adenocarcinoma	1.57
lung adenocarcinoma	1.38
excision of lung	1.36
suspected lung cancer	1.36
smoking	1.35
histological type	1.34
staging	1.3
incision of lung	1.25
radon	1.23
squamous carcinoma	1.22
stage iiia	1.20

(a)

Concept	Weight
nonsmall cell lung cancer	2.77
small cell lung cancer	2.15
lung carcinoma	1.64
excision of lung	1.36
incision of lung	1.25
bronchial carcinomas	1.20
lung tumor	1.14
mediastinoscopy	1.08
pneumonectomy	0.99
resection of trachea	0.95
lung cancer screening	0.89
thoracotomy	0.87
lung collapse therapy	0.84
percutaneous cordotomy	0.71
lung cancer prevention	0.71
remote afterloaders	0.70
stages microscope	0.69
neoadjuvant therapy	0.68
lobectomy	0.66
beta carotene	0.63

(b)

Concept	Weight
nonsmall cell lung cancer	2.77
small cell lung cancer	2.15
lung carcinoma	1.64
bronchial carcinomas	1.2
lung tumor	1.14
mediastinoscopy	1.08
thoracotomy	0.87
ipomeanol	0.86
repairmen	0.83
mediast lymph nodes sampling	0.79
carotenoid	0.7
beta carotene	0.63
pleura cancer	0.62
fiberoptic bronchoscopy with biopsy	0.59
mediastinal metastasis	0.57
chest x ray	0.57
hospital porter	0.57
platinol	0.57
diagnosis	0.57
staging	0.55

(c)

Figure 4: (a) Statistical expansion concepts for query "lung cancer, *treatment*." (b) Knowledge-based expansion concepts for query "lung cancer, *treatment*." (c) Knowledge-based expansion concepts for query "lung cancer, *diagnosis*."

cepts) to the expanded query. The surrounding concepts for "lung cancer," for example, are illustrated in Figure 3.

In our study, we have also tried expanding more than the immediate surrounding concepts, e.g., ancestors or descendants more than two levels from c_k. Our results reveal that enlarging the scope of surrounding concepts yields degraded performance, which is consistent with results reported by Hersh et al. [27]. As a result, in our experiments, we restrict the scope to parents, children and siblings of c_k only.

For illustration purposes, for the sample query "lung cancer, treatment," we first use statistical expansion technique to derive candidate expansion concepts, and then identify the scenario-specific expansion concepts using the procedure described above. The top-20 heavily-weighted statistical expansion concepts are listed in Figure 4(a), where the weights are assigned according to the co-occurrence thesaurus (Section 3.1). The shaded concepts in Figure 4(a) are the ones identified as scenario-specific, corresponding to the concepts in the shaded circles of Figure 2. These scenario-specific concepts, together with other top-weighted scenario-specific concepts, are shown in Figure 4(b). Some concepts down the list of Figure 4(b) (e.g., "lung collapse therapy") do not appear in the list of Figure 4(a), simply because they have relatively smaller weights and we are only showing the top-20 statistically-related concepts in Figure 4(a).

Similar to Figure 4(b), we have also derived scenario-specific expansion concepts for another query "lung cancer, diagnosis," and show results in Figure 4(c). The following observations are made from these results.

- By comparing Figure 4(a) with Figure 4(b), we can clearly see that knowledge-based expansion identifies expansion concepts that are much more relevant to the original query's scenario ("treatment") compared to statistical expansion.
- By comparing Figure 4(b) with Figure 4(c), we can see that the results of knowledge-based expansion differ under different scenarios, i.e., "treatment" and "diagnosis," thus achieving the goal of scenario-specific query expansion.

Adjusting the weights of the scenario-specific expansion concepts to increase their significance. By comparing the weights of scenario-specific expansion concepts with those of statistical expansion concepts (e.g., comparing the weights in Figure 4(b) with those in Figure 4(a)), we can see that scenario-specific concepts generally have less weights. This happens because we have filtered out certain heavily-weighted concepts, concepts that are statistically-related but not scenario-specific. Because of their relatively smaller weights, the scenario-specific concepts appended by the knowledge-based method bring less impact to the expanded query, compared to that in the statistical method.

To increase the impact of the scenario-specific concepts, we can "boost" their weights by multiplying a linear factor, so that the overall "significance" of the scenario-specific concepts is comparable to that of the statistical-expansion concepts. To quantify the "significance" of a set of expansion concepts, we use the length of the expansion vector composed by these concepts. Formally, let $|V|$ represent the length of a l-dimension vector $V = (v_1, v_2, ..., v_l)$, where $|V|$ is computed as:

$$|V| = \sqrt{v_1^2 + v_2^2 + \cdots + v_l^2}$$

Further, let V_{stat} represent the vector of statistical expansion concepts and V_{KB} represent the vector of scenario-specific expansion concepts generated by the knowledge-based method. Because certain heavily weighted components in V_{stat} has been filtered out to generate V_{KB}, for any query we have:

$$|V_{stat}| \geq |V_{KB}|$$

We define the *boosting factor* for V_{KB} to be:

$$1 + \alpha \cdot (\frac{|V_{stat}|}{|V_{KB}|} - 1) \quad (1)$$

Here α is a positive real number that controls the length of the scenario-specific expansion vector after boosting. If $\alpha = 0$, the boosting factor is reduced to 1 which essentially means no boosting; If $\alpha = 1$, the boosting factor is reduced to $\frac{|V_{stat}|}{|V_{KB}|}$ which makes the boosted vector have exactly the same length as that of the statistical expansion vector.

In the experiments section, we will discuss how this parameter α affects the retrieval result.

3.3 Knowledge acquisition

The quality of our knowledge-based method described in Section 3.2 is largely dependent on the quality and completeness of the domain-specific knowledge source. The knowledge source used in our study, UMLS, is not specifically designed for the purpose of scenario-specific retrieval. As a result, in our study we have observed some frequently-asked scenarios (e.g. query scenarios in OHSUMED [20]) that are undefined in UMLS. To support these scenarios, we propose the following methodology for knowledge acquisition to supplement the UMLS knowledge source.

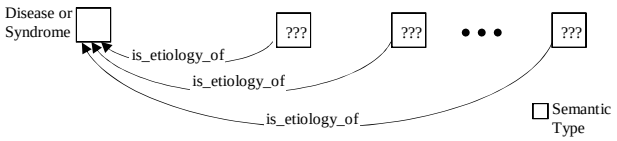


Figure 5: Supplementing the UMLS Semantic Network with relationship graphs for previously undefined scenarios

First, we identify the scenarios that are not currently supported by UMLS. By studying sample medical queries, e.g., queries in the OHSUMED test collection, we have identified the following list of scenarios that are frequently used but undefined by UMLS: "differential diagnosis," "etiology," "risk factors," "pathophysiology," "prognosis," "epidemiology," "research," "organisms" of a disease, "complications" of a disease/medication, "criteria" of or "when to perform" a medication, and "preventive health care" for a type of patients. By "undefined," we mean that such scenarios have no corresponding relationship graphs in the UMLS Semantic Network, such as the graph shown in Figure 2. Therefore, we plan to supplement the UMLS Semantic Network with additional relationship graphs to support the above frequently-used scenarios.

We use the following method for this supplementation task: First, we present to medical experts a blank Semantic Type relationship graph such as the one shown in Figure 5. Edges in this relationship graph are labelled with one of the undefined scenario, e.g., "etiology." The experts will decide which UMLS Semantic Types should be filled into the blank rectangles. (Currently UMLS defines 134 Semantic Types.) For example, because viruses are related to the etiology of a wide variety of diseases, the Semantic Type "Virus" will be filled into one of the rectangles in Figure 5. Note that the number of black rectangles are not pre-determined and will be decided by the experts to make sure the relationship graph is complete.

4. EXPERIMENTAL RESULTS

4.1 Dataset and experimental setup

Dataset. Our experiment is based on the OHSUMED [20] test collection that has been widely used in medical-information-retrieval research. OHSUMED consists of 1) a corpus, 2) a query set, and 3) relevance judgments for each query.

- Corpus. The corpus consists of 348,000 MEDLINE articles from 1988 to 1992. Each document contains a title, an optional abstract, a set of MeSH headings, author information, publication type, source, a MEDLINE identifier, and a document ID.
- Query set. The query set consists of 106 queries. Each query contains a patient description, an information request, and a query ID. Since we are interested in short and general queries, we use the information-request sub-portion to represent each query. To study scenario-specific retrieval, we focus on all queries in the form of "⟨key concept(s)⟩, ⟨scenario concept(s)⟩." Among the 106 queries, 57 queries satisfy this criterion and are included in our study.[4] The rest of the queries skipped in our study typically ask for the relationship among several key concepts without mentioning scenario concepts, e.g., "use of beta-blockers for thyrotoxicosis during pregnancy" or "chemotherapy advanced for advanced metastatic breast cancer."

[4]In fact there is an additional query, query #8, which also satisfy this criteria. However, OHSUMED provides no relevance judgements for this query, and therefore we exclude this query from our experiments.

Scenario	Queries ID's
treatment of a disease	2, 13, 15, 16, 27, 29, 30, 31, 32, 35, 37, 38, 39, 40, 42, 43, 45, 53, 56, 57, 58, 62, 67, 69, 72, 74, 75, 76, 77, 79, 81, 85, 93, 98, 102
diagnosis of a disease	15, 21, 37, 53, 57, 58, 72, 80, 81, 82, 97
prevention of a disease	64, 85
differential diagnosis of a symptom/disease	14, 23, 41, 43, 47, 51, 65, 69, 70, 74, 76, 103
pathophysiology of a disease	2, 3, 26, 64, 77
complications of a disease/medication	3, 30, 52, 61, 62, 66, 79
etiology of a disease	14, 26, 29
risk factors of a disease	35, 64, 85
prognosis of a disease	45
epidemiology of a disease	3
research of a disease	75
organisms of a disease	81
criteria of medication	49, 52, 94
when to perform a medication	33
preventive health care for a type of patients	96

Figure 6: Number of queries mentioning each scenario

Scenario	# of relevant Semantic Types in the relation-ship graph for that scenario
differential diagnosis of a symptom/disease	10
pathophysiology of a disease	56
complications of a disease/medication	15
etiology of a disease	40
risk factors of a disease	40
prognosis of a disease	15
epidemiology of a disease	13
research of a disease	28
organisms of a disease	7
criteria of medication	26
when to perform a medication	5
preventive health care for a type of patients	10

Figure 7: Number of Semantic Types in the relationship graphs after knowledge acquisition

We list the queries that mention each scenario in Figure 6. Due to space constraint, we only provide the query ID's. The query strings can be downloaded at OHSUMED's official Website.[5] Note that some queries do not mention the scenario terms such as "treatment" or "diagnosis" directly, but "management" or "workup" instead. We consult experts in UCLA Medical School to classify these queries into the appropriate scenarios.

- Relevance judgements. For a given OHSUMED query, a document is either judged by experts as definitely-relevant (DR), partially-relevant (PR), irrelevant or not judged at all. In our experiments, we restrict the retrieval to the 14,430 judged documents only and count both DR and PR documents as relevant answers as we measure the precision-recall of a particular retrieval method.[6]

Indexing and VSM. We index both documents and queries using word stems, and assign weight to each stem using the standard $tf \cdot idf$ weighting scheme [6]. Word stems are derived using the Lovins stemmer [28]. Special considerations in this indexing process include:

- We use the title and the abstract to index each document. We have discarded the MeSH headings in indexing in order to simulate a common application environment in which no

[5]`ftp://medir.ohsu.edu/pub/OHSUMED`

[6]Treating both DR and PR documents are relevant documents is consistent with the settings of existing studies [20, 27]

expert-assigned indexing terms are available.

- To emphasize the importance of title terms in representing a document's content, we count the tf of every single appearance of a term in the title as 3, while keeping the tf for terms in other parts of a document unmodified.
- Since the expanded query is eventually represented as a vector of stems, we use the following procedure to convert the expansion concepts (derived either by our knowledge-based method or the statistical method) to word stems and append these stems to the original query: For each expansion concept we first look up its concepts strings from UMLS. We further remove all stop words from these concept strings and convert all the words into word stems. The weights of these expansion stems are assigned based on the co-occurrence thesaurus computed from the corpus [13].

After we index the documents and the expanded query using word stems, we use the standard stem-based Vector Space Model (VSM) [6] to compute query-document similarities and generate document ranking.

4.2 Knowledge acquisition results

We follow the methodology in Section 3.3 for this task. To supplement the Semantic Network with additional relationship graphs for the currently unsupported scenarios (e.g. "etiology" of a disease), we interviewed two medical experts at UCLA Medical School. During the interview we first described the meaning of relationship graphs such as Figure 5, and then presented the entire list of UMLS Semantic Types to the experts so that appropriate Semantic Types were filled into the question marks. We communicated the results by one expert with another until they reached a consensus. Basic statistics for the knowledge acquired in this step are presented in Figure 7. The detailed list of Semantic Types for each scenario is presented in the extended version of this paper [29].

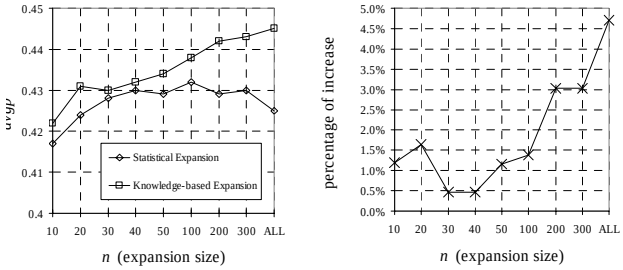
4.3 Retrieval results

In this section we study the performance of knowledge-based expansion compared to that of statistical expansion. We first compare the two methods under different expansion sizes, then study the performance of the knowledge-based method under different boosting factors and different query scenarios.

4.3.1 Comparison of the two methods under different expansion sizes

For a given expansion size n, we use both knowledge-based expansion and statistical expansion to expand the top-n stems that have the heaviest weights. For knowledge-based expansion, no weight boosting is applied at this stage. We measure the performance of both methods using the 11-point precision average, denoted as *avgp*. We have also compared the two methods using other metrics, such as precision among the top-10 or top-20 retrieved documents, and the comparison results are similar.

We compute *avgp* for both methods on each of the 57 queries, and further average the results over the 57 queries. Figure 8(a) shows the performance of the two methods, whereas Figure 8(b) shows the percentage of improvement of knowledge-based expansion over statistical expansion. In these figures, "n=All" means appending all expansion terms that have non-zero weights into the original query. Before we compare the results, we emphasize that the baseline method in our comparison, the statistical expansion method, outperforms the no-expansion retrieval method by more than 5% under most of the settings. (The *avgp* for no-expansion retrieval is 0.408.)



(a) *avgp* of the two methods (b) percentage of improvement

Figure 8: Comparison of the two methods using *avgp*

α	n: 10	20	30	40	50	100	200	300	All
0 (no boosting)	0.422 (1.2%)	0.431 (1.7%)	0.430 (0.5%)	0.432 (0.5%)	0.434 (1.2%)	0.438 (1.4%)	0.442 (3.0%)	0.443 (3.0%)	0.445 (4.7%)
0.25	0.425 (1.9%)	0.436 (2.8%)	0.435 (1.6%)	0.436 (1.4%)	0.439 (2.3%)	0.443 (2.5%)	0.445 (3.7%)	0.448 (4.2%)	0.450 (5.9%)
0.5	0.426 (2.2%)	0.435 (2.6%)	0.438 (2.3%)	0.439 (2.1%)	0.440 (2.6%)	0.444 (2.8%)	0.447 (4.2%)	0.450 (4.7%)	0.451 (6.1%)
0.75	0.428 (2.6%)	0.436 (2.8%)	0.437 (2.1%)	0.439 (2.1%)	0.440 (2.6%)	0.444 (2.8%)	0.447 (4.2%)	0.450 (4.7%)	0.450 (5.9%)
1	0.428 (2.6%)	0.436 (2.8%)	0.437 (2.1%)	0.437 (1.6%)	0.439 (2.3%)	0.443 (2.5%)	0.446 (4%)	0.450 (4.7%)	***0.452 (6.4%)***
1.25	0.426 (2.2%)	0.436 (2.8%)	0.435 (1.6%)	0.437 (1.6%)	0.439 (2.3%)	0.442 (2.3%)	0.445 (3.7%)	0.449 (4.4%)	0.450 (5.9%)
1.5	0.425 (1.9%)	0.435 (2.6%)	0.434 (1.4%)	0.436 (1.4%)	0.439 (2.3%)	0.439 (1.6%)	0.443 (3.3%)	0.445 (3.5%)	0.447 (5.2%)

Figure 9: The impact of different weight-boosting settings on the performance of knowledge-based expansion

As the figure shows, the performance for knowledge-based expansion generally increases as n increases, and usually reaches the peak when n=All. On the other hand, the performance of the statistical method degrades after n=100. This is due to the fact that statistical expansion does not distinguish between expansion terms that are scenario-specific from those that are not. As a consequence, as more terms are appended to the original query, the negative impact of those non-scenario-specific terms begins to accumulate and after a certain point the performance drops. In contrast, the knowledge-based method appends scenario-specific terms only, and consequently, the performance of the knowledge-based method keeps increasing as more "useful" terms are appended.

4.3.2 The impact of weight boosting on the performance of knowledge-based expansion

In the next experiments, we multiply a boosting factor to the weights of knowledge-based expansion terms. The boosting factor is computed using Eq. 1, under the different settings of $\alpha = 0.25, 0.5, 0.75, 1, 1.25, 1.5$. Figure 9 shows the impact of different boosting amount on the performance of knowledge-based expansion. Each cell in the figure shows 1) the performance of knowledge-based expansion and 2) the improvement of knowledge-based expansion over statistical expansion under the same expansion size. Thick-bordered cells represent the best performance within each column (i.e. under the same setting of expansion size); Shaded cells represent the best performance within each row (i.e. under the same setting of boosting factor). The best performance in the entire figure is highlighted in bold and italic.

The following observation can be made from these results:

- Boosting helps improve the performance of knowledge-based expansion, under all expansion sizes. Settings $\alpha = 0.5$ or $= 0.75$ generally yield the best boosting effect.
- Given a fixed boosting setting, having a larger expansion size n helps improve the performance. The best performance under all α settings is consistently achieved by setting n=All.

		scenario				
		treatment of a disease	differential diagnosis of a symptom / disease	diagnosis of a disease	complication of a disease / medication	pathophysiology of a disease
α	0 (no boosting)	0.465 (3.9%)	0.444 (9.4%)	0.464 (7.5%)	0.466 (2.4%)	0.564 (0.5%)
	0.25	0.470 (5.2%)	0.444 (9.4%)	0.470 (9.0%)	0.470 (3.1%)	0.569 (1.4%)
	0.5	0.474 (5.9%)	0.439 (8.0%)	0.472 (9.4%)	0.470 (3.2%)	0.571 (1.8%)
	0.75	0.474 (6.0%)	0.434 (6.8%)	0.473 (9.7%)	0.464 (2.0%)	0.573 (2.3%)
	1	0.474 (5.9%)	0.438 (7.9%)	0.474 (9.8%)	0.466 (2.4%)	0.580 (3.4%)
	1.25	0.472 (5.4%)	0.433 (6.6%)	0.480 (11%)	0.470 (3.1%)	0.579 (3.3%)
	1.5	0.466 (4.2%)	0.431 (6.1%)	0.475 (9.9%)	0.467 (2.6%)	0.579 (3.3%)

Figure 10: The performance of knowledge-based expansion in different scenarios. Expansion size n=All

4.3.3 *Performance of knowledge-based expansion in different scenarios*

In our next experiment, we study how knowledge-based expansion perform in different scenarios. To do this, we group the 57 queries according to the scenarios they ask for, and we select the largest five groups, namely "treatment," "diagnosis," "pathophysiology" of a disease, "differential diagnosis" of a symptom/disease and "complications" of a disease/medication. We skip the rest of the scenarios because each of these scenarios has too few number of queries to derive reliable statistics. (The number of queries that ask for each scenario is shown in Figure 6.)

We further average the performance of knowledge-based expansion within each group of queries, and show the *avgp* results in Figure 10. Similar to the previous figure, each cell shows 1) the performance of knowledge-based expansion averaged over the corresponding group of queries, and 2) the improvement of knowledge-based expansion over statistical expansion under the same settings. For example, the shaded cell in Figure 10 shows that, among the 35 queries that ask about the "treatment" scenario and under the boosting setting of $\alpha = 0.75$, knowledge-based expansion achieves an average *avgp* of 0.474. This represents a 6.0% improvement over the statistical method measured within the same group of queries.

To derive the results in Figure 10, we set the expansion size n=All which allows the knowledge-based method to yield the best performance.

These results generally suggest that knowledge-based expansion performs differently for queries with different scenarios. More specifically, the method yields more improvements in the "treatment," "differential diagnosis" and "diagnosis" scenarios. In contrast, it yields less improvements in the "complication" and "pathophysiology" scenarios. A possible explanation lies in the different knowledge structures for these five scenarios. In the relationship graphs defined for the latter two scenarios (i.e. "complication" and "pathophysiology"), there are more relevant Semantic Types than those in the former three scenarios (Figure 7). As a consequence, when handling queries with the latter two scenarios, the knowledge-based method keeps more concepts as scenario-specific expansion concepts during the filtering step. Thus the expansion result of the knowledge-based method resembles that of the statistical expansion method, leading to close performance between the two methods.

5. RELATED WORKS

Query expansion, as an effective method to ameliorate the query-document mismatch problem, has been studied for decades. An overview of various query expansion techniques can be found in [8]. The basic idea behind all techniques is to supplement the original query with additional terms related to the original query topic, so that the modified query has a better chance to match relevant documents. The following specific techniques, in a broader sense, fall underneath the general umbrella of query expansion.

- *Manual expansion.* A human expert or the user manually looks at the original query and selects from a knowledge source (e.g. WordNet) the best terms to expand [30, 31].
- *Relevance feedback.* The expansion terms are selected from a few top-ranked documents that are manually marked by the user as relevant answers [25, 26]. In certain cases, terms from those documents marked as irrelevant will also be "subtracted" from the original query.
- *Statistical expansion (or automatic expansion).* The expansion terms are automatically selected either from a term co-occurrence thesaurus [13, 14, 15, 32] or pseudo-relevance feedback results [21, 22, 23, 24, 32, 16].

These past research efforts do not attempt to automatically exploit a domain-specific knowledge source to refine the query expansion results and provide scenario-specific expansion.

Recently with the emergence of UMLS, a full-fledged knowledge source in the medical domain, methods have been proposed to automatically utilize this knowledge source in query expansion. Aronson et al. [33] proposed to use MetaMap [17], a program that maps medical free text to UMLS concepts, to first identify concepts mentioned by the original query. Their approach further expands synonyms of the original query concepts, with the guidance of UMLS. Hersh et al. [27] proposed to expand the parent and child concepts of the original query concepts, based on the concept hierarchy defined in the UMLS Metathesaurus (e.g. Figure 3). Our research differs from these works in the following aspects:

- Our research targets one type of medical queries, namely scenario-specific queries, that have been shown to be predominant among medical users' search requests [1, 2, 3, 4, 5]. In dealing with such queries, it is often too narrow to expand just the synonyms or parent/child concepts without considering the scenario information embedded in the original query. For example, previous methods will exclude "lung excision" from the expansion list for query "lung cancer, treatment," simply because "lung excision" is neither a synonym nor a parent/child concept of any original query concept.

  In contrast, our method explores the scenario information in the original query, relates that information to certain knowledge structures in UMLS (more specifically, the UMLS Semantic Network) and uses the identified knowledge structure to guide the selection of scenario-specific concepts. The resulting expansion will have a much broader scope than just synonyms and/or parent/child concepts.
- Previous works only compare against a baseline generated by no query expansion. To the best of our knowledge, we are the first to compare against statistical expansion. Since statistical expansion has also been shown to be effective in improving retrieval performance [13, 14, 15, 21, 22, 23, 24, 32, 16], it is crucial to make the second type of comparison in order to study the true impact of a knowledge source in query expansion. (In our experiments we also observe that statistical expansion outperforms the no-expansion method by at least 5% in most of the cases.) Our study shows that even when comparing with statistical expansion, the knowledge-based method yields reasonable improvements.

  In fact, the same dataset (OHSUMED) has been used in both our study and that of Hersh et al. [27]. However, Hersh et al. reported degrading performance by their query expansion approach compared to the no-expansion method. We attribute the differences between our results and theirs to two factors: 1) We study a subset of OHSUMED queries that are

scenario-specific; 2) We apply a knowledge-based method that is designed to effectively handle such scenario-specific queries.

6. CONCLUSION

Scenario-specific queries represent a special type of queries that frequently appear in medical free-text retrieval. In this research, we have proposed a knowledge-based query expansion method to improve the retrieval performance for such queries. More specifically, the contributions of this work are the following:

- We have designed a method that automatically exploits the knowledge structures in the UMLS Semantic Network and the UMLS Metathesaurus to identify concepts that are specifically related to the scenario(s) in the original query. Appending such identified concepts to the original query results in scenario-specific expansion.
- Given that a knowledge-source is usually incomplete in handling all scenarios appearing in real queries, we have proposed a methodology to supplement the knowledge source.
- We have performed extensive experimental evaluation of the knowledge-based method by comparing against the statistical expansion method. Our experimental study has shown that:
  - Our proposed knowledge-based method is able to create scenario-specific query expansion, and yields improvements over statistical expansion when handling scenario-specific queries.
  - Since knowledge-based expansion tends to expand terms with smaller weights into the original query, boosting the weights of these terms is necessary to generate reasonable improvements over the statistical method.
  - The knowledge-based expansion method performs differently for different query scenarios. This happens because the knowledge structures defined for these scenarios exhibit different characteristics.

7. ACKNOWLEDGEMENTS

This research is supported in part by NIC/NIH Grant #4442511-33780. We are grateful to Andrew Chen and Wei Liu from UCLA Medical School for providing their domain knowledge when we performed the knowledge acquisition task (Section 3.3), and we also thank other researchers in UCLA Department of Radiology Sciences who are supported by the same NIC/NIH grant for valuable discussion and feedbacks.

8. REFERENCES

[1] R. Haynes, K. McKibbon, C. Walker, N. Ryan, D. Fitzgerald, and M. Ramsden. Online access to medline in clinical settings. *Ann Intern Med*, 112:78–84, 1990.

[2] W.R. Hersh, J. Pentecost, and D.H. Hickam. A task-oriented approach to information retrieval evaluation. *JASIS*, 47(1):50–56, 1996.

[3] J.W. Ely, J.A. Osheroff, M.H. Ebell, G.R. Bergus, B.T. Levy, M.L. Chambliss, and E.R. Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7):211–220, 1999.

[4] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, and P.Z. Stavri. A taxonomy of generic clinical questions: classification study. *BMJ*, 321(12):429–432, 2000.

[5] N.L. Wilczynski, K.A. McKibbon, and R.B. Haynes. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *International Journal of Medical Informatics*, 10(1):390–393, 2001.

[6] G. Salton and M.J. McGill. *Introduction to Mordern Information Retrieval*. McGraw Hill, 1983.

[7] J.P. Callan, W.B. Croft, and S.M. Harding. The INQUERY retrieval system. In *Proceedings of DEXA '92*, 1992.

[8] E.N. Efthimiadis. Query expansion. *Annual Review of Information Science and Technology*, 31:121–187, 1996.

[9] D.G. Covell, G.C. Uman, and P.R. Manning. Information needs in office practice: are they being met? *Ann Intern Med*, 103:596–599, 1985.

[10] J. Marshall. The continuation of end-user on line searching by health professionals: preliminary survey results. In *Proceedings of the Medical Library Association Annual Meeting*, 1990.

[11] P.N. Gorman and M. Helfand. Information seeking in primary care: How physicians choose which clinical questions to pursue and which to leave unanswered. *Med Decis Making*, 15(2):113–119, 1995.

[12] P.N. Gorman, J. Ash, and L. Wykoff. Can primary care physicians' questions be answered using the medical literature? *Bull Med Lib Assoc*, 82:140–146, 1994.

[13] Y. Qiu and H.P. Frei. Concept-based query expansion. In *Proceedings of ACM SIGIR '93*, 1993.

[14] Y. Jing and W.B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO '94*, 1994.

[15] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In *Proceedings of ACM SIGIR '96*, 1996.

[16] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of ACM SIGIR '98*, 1998.

[17] A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceedings of AMIA Annual Symp 2001*, 2001.

[18] Q. Zou, W.W. Chu, C. Morioka, G.H. Leazer, and H. Kangarloo. IndexFinder: A method of extracting key concepts from clinincal texts for indexing. In *Proceedings of AMIA Annual Symp 2003*, 2003.

[19] National Library of Medicine. *UMLS Knowledge Sources*. 12^{th} edition, 2001.

[20] W. Hersh, C. Buckley, T.J. Leone, and D. Hickam. OHSUMED: an insteractive retrival evaluation and new large test collection for research. In *Proceedings of ACM SIGIR '94*, 1994.

[21] E.N. Efthimiadis and P. Biron. UCLA-okapi at TREC-2: Query expansion experiments. In *Proceedings of the Second Text REtrieval Conference (TREC-2)*, 1993.

[22] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1994.

[23] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1994.

[24] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New retrieval approaches using SMART: TREC-4. In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, 1995.

[25] J.J. Rocchio. *The SMART Retrieval System - Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval. Prentice Hall, 1971.

[26] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[27] W.H. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the UMLS metathesaurus. In *Proceedings of AMIA Annual Symp 2000*, 2000.

[28] J.B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1):11–31, 1968.

[29] Z. Liu and W.W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Technical report, Computer Science Department, UCLA, 2004.

[30] E.M. Voorhees. On expanding query vectors with lexically related words. In *Proceedings of TREC-2*, pages 223–232, 1993.

[31] E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of ACM SIGIR '94*, pages 61–69, 1994.

[32] P. Srinivasan. Query expansion and MEDLINE. *Information Processing and Management*, 32(4):431–443, 1996.

[33] A.R. Aronson and T.C. Rindflesch. Query expansion using the UMLS. In *Proceedings of AMIA Annual Symp 1997*, 1997.