Knowledge-Based Query Expansion to Support Scenario-Specific Retrieval of Medical Free Text[1]

Zhenyu Liu
UCLA Computer Science Department
Los Angeles, CA 90024
vicliu@cs.ucla.edu

Wesley W. Chu
UCLA Computer Science Department
Los Angeles, CA 90024
wwc@cs.ucla.edu

Abstract

In retrieving medical free text, users are often interested in answers pertinent to certain scenarios that correspond to common tasks performed in medical practice, e.g., treatment or diagnosis of a disease. A major challenge in handling such queries is that scenario terms in the query (e.g. treatment) are often too general to match specialized terms in relevant documents (e.g. chemotherapy). 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 compared the proposed method with traditional statistical expansion that expands terms which are statistically correlated but not necessarily scenario specific. Our study on two standard testbeds shows that the knowledge-based method, by providing scenario-specific expansion, yields notable improvements over the statistical method in terms of average precision-recall. On the OHSUMED testbed, for example, the improvement is more than 5% averaging over all scenario-specific queries studied and about 10% for queries that mention certain scenarios, such as treatment of a disease and differential diagnosis of a symptom/disease.

1 Introduction

In recent years, there has been a phenomenal growth of online medical document collections. Collections such as PubMed[2] and MedlinePlus[3] provide comprehensive coverage of medical literature and teaching materials. In searching these collections, it

[1] This research is supported in part by NIC/NIH Grant #4442511-33780

[2] http://www.pubmed.gov/

[3] http://www.medlineplus.com

is desirable to retrieve only those documents pertaining to a specific medical "scenario," where a scenario is defined as a frequently-reappearing medical task. For example, in treating a lung cancer patient, a physician may pose the query `lung cancer treatment` in order to find the latest treatment techniques for this disease. Here, `treatment` is the medical task that marks the scenario for this query. Recent studies [Haynes et al.(1990), Hersh et al.(1996), Ely et al.(1999), Ely et al.(2000), Wilczynski et al.(2001)] reveal that in clinical practice, as many as 60% of physicians' queries center on 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 [Salton and McGill(1983)] or INQUIRY [Callan et al.(1992)]). Such systems suffer from the fundamental problem of *query-document mismatch* [Efthimiadis(1996)] when handling scenario-specific queries. Scenario terms in these queries are represented using general terms, e.g., the term `treatment` in the query `lung cancer treatment`. On the contrary, in full-text medical documents, more specialized terms such as `lung excision` or `chemotherapy` are used to express the same topic. Such mismatch of terms leads to poor retrieval performance [Zeng et al.(2002), Tse and Soergel(2003)].

There has been a substantial amount of research on *query expansion* [Qiu and Frei(1993), Jing and Croft(1994), Buckley et al.(1994), Robertson et al.(1994), Buckley et al.(1995), Xu and Croft(1996), Srinivasan(1996), Mitra et al.(1998)] that ameliorates the query-document mismatch problem. However, such techniques also have difficulties handling scenario-specific queries. Query expansion appends 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.

In the domain of medical text retrieval, researchers have proposed to exploit the *Unified Medical Language System (UMLS)*, a full-fledged knowledge source in the medical domain, to expand the original query with related terms and to improve retrieval performance. Current approaches either explore the synonym relationships defined in UMLS and expands synonyms of the original query terms [Aronson and Rindflesch(1997), Plovnick and Zeng(2004), Guo et al.(2004)] or explore the hypernym/hyponym relationships and expands terms that have wider/narrower meaning than the original query terms [Hersh et al.(2000)]. Extensive evaluation of these approaches has been performed on standard testbeds such as OHSUMED [Aronson and Rindflesch(1997), Hersh et al.(2000)] and the TREC Genomics ad hoc topics [Guo et al.(2004)]. However, no study has consistently produced significant differences in retrieval effectiveness before and after expansion. Particularly, we note that when handling scenario-specific queries, such solutions still generally suffer from the query-document mismatch problem. For example, the synonyms, hypernyms or hyponyms for all the terms in query `lung cancer treatment`, as defined by the knowledge source, are `lung carcinoma`, `cancer`, `therapy`, `medical procedure`, etc. With such terms expanded, the query will still have difficulty matching documents that extensively use specialized terms such as `chemotherapy` and `lung excision`.

In this paper, we propose a *knowledge-based query expansion* technique to support *scenario-specific* retrieval. Our technique exploits domain knowledge to restrict query expansion to scenario-specific terms and yields better retrieval performance than that of traditional query expansion approaches. The following are challenges in developing such a knowledge-based technique:

- **Using domain knowledge to automatically identify scenario-specific terms.** It is impractical to ask users or domain experts to manually identify scenario-specific terms for every query and all possible scenarios. 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 treat this distinction, we propose to exploit a domain-specific knowledge source.
- **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. 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 organized as follows. We first present a framework for knowledge-based query expansion in Section 2. We then describe the detailed method in this framework in Section 3. We experimentally evaluate the method and report the results in Section 4. In Section 5, we address the issue of supplementing a knowledge source via knowledge acquisition. We further discuss the relevancy of expansion terms judged by domain experts in Section 6.

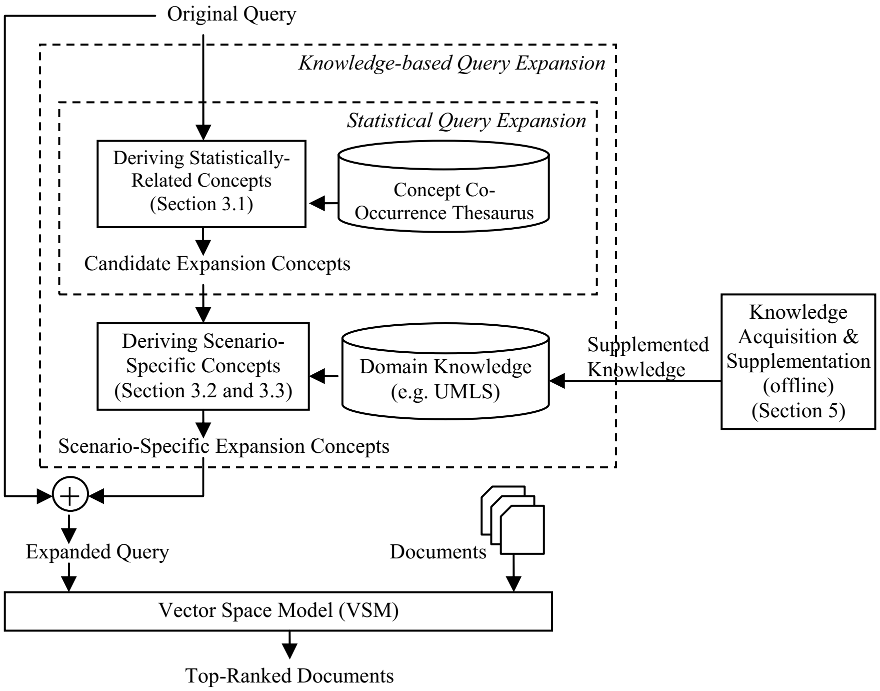

Original Query
Knowledge-based Query Expansion
Statistical Query Expansion
Deriving Statistically-Related Concepts (Section 3.1)
Concept Co-Occurrence Thesaurus
Candidate Expansion Concepts
Deriving Scenario-Specific Concepts (Section 3.2 and 3.3)
Domain Knowledge (e.g. UMLS)
Supplemented Knowledge
Knowledge Acquisition & Supplementation (offline) (Section 5)
Scenario-Specific Expansion Concepts
Expanded Query
Documents
Vector Space Model (VSM)
Top-Ranked Documents

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

2 A Framework for Knowledge-Based Query Expansion

Figure 1 depicts the components in a knowledge-based query expansion and retrieval framework. For a given query, *Statistical Query Expansion* (whose scope is marked by the inner dotted rectangle) will first derive *candidate expansion concepts*[4] that are statistically co-occurring with the given 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 a domain knowledge source such as UMLS [NLM(2001)] (Section 3.2). Such knowledge may be incomplete and fail to include all possible query scenarios. Therefore, in an off-line process, we apply a *Knowledge Acquisition and Supplementation* module to supplement the incomplete knowledge (Section 5).

[4] 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 [Aronson(2001)], IndexFinder [Zou et al.(2003)], etc.

After the query is expanded with scenario-specific concepts, we employ a *Vector Space Model* (VSM) to compare the similarity between the expanded query and each document. Top-ranked documents with the highest similarity measures are output to the user.

3 Method

Formally, the problem for knowledge-based query expansion can be stated as follows: Given a scenario-specific query with a key concept denoted as c_{key} (e.g., `lung cancer` or `keratoconus`[5]) and a set of scenario concepts denoted as c_s (e.g., `treatment` or `diagnosis`), we need to derive specialized concepts that are related to c_{key} and the relations should be specific to the scenarios defined by c_s.

In this section, we describe how to derive such scenario-specific concepts first by presenting existing statistical query expansion methods which generate candidates for such scenario-specific concepts. We then propose a knowledge-based method that selects scenario-specific concepts from this candidate set with the aid of a domain knowledge source.

3.1 Deriving Statistically-Related Expansion Concepts

Statistical expansion is also referred to as *automatic query expansion* [Efthimiadis(1996), Mitra et al.(1998)]. The basic idea is to derive concepts that are statistically related to the given query concepts, where the statistical correlation is derived from a document collection (e.g., OHSUMED [Hersh et al.(1994)]). Appending such concepts to the original query makes the query expression more specialized and helps the query better match relevant documents. Depending on how such statistically-related concepts are derived, statistical expansion methods fall into two major categories:

- *Co-occurrence-thesaurus-based expansion* [Qiu and Frei(1993), Jing and Croft(1994), Xu and Croft(1996)]. 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 matrix. A detailed procedure for computing the co-occurrence matrix and for assigning weights to expansion concepts can be found in [Qiu and Frei(1993)].
- *Pseudo-relevance-feedback-based expansion* [Efthimiadis and Biron(1993), Buckley et al.(1994), Robertson et al.(1994), Buckley et al.(1995),

[5] An eye disease

#	Concepts that statistically correlate to keratoconus
1	fuchs dystrophy
2	penetrating keratoplasty
3	epikeratoplasty
4	corneal ectasia
5	acute hydrops
6	keratometry
7	corneal topography
8	corneal
9	aphakic corneal edema
10	epikeratophakia
11	granular dystrophy corneal
12	keratoplasty
13	central cornea
14	contact lens
15	ghost vessels

Table 1: Concepts that statistically correlate to `keratoconus`

#	Concepts that *treat* keratoconus
1	penetrating keratoplasty
2	epikeratoplasty
3	epikeratophakia
4	keratoplasty
5	contact lens
6	thermokeratoplasty
7	button
8	secondary lens implant
9	fittings adapters
10	esthesiometer
11	griffonia
12	trephine
13	slit lamps
14	fistulization
15	soft contact lens

(a)

Concepts that *diagnose* keratoconus
keratometry
corneal topography
slit lamp examination
topical corticosteroid
echocardiography 2 d
tem
interferon
alferon
analysis
microscopy
bleb
tetanus toxoid
antineoplastic
heart auscultation
chlorbutin

(b)

Table 2: Concepts that `treat` or `diagnose` `keratoconus`

> Mitra et al.(1998)]. 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 except that, instead of asking users to identify relevant documents as feedback, top-ranked (e.g. top-10) documents are automatically treated as "pseudo" relevant documents and are inserted into the feedback loop. Weight assignment in pseudo relevance feedback [Buckley et al.(1994)] typically follows the same weighting scheme ($\langle \alpha, \beta, \gamma \rangle$) for conventional relevance feedback techniques [Rocchio(1971)].

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 used the co-occurrence-thesaurus-based method to derive statistically-related concepts. For convenience of discussion, we use $co(c_i, c_j)$ to denote the co-occurrence between concept c_i and c_j, a value that appears as the $\langle i, j \rangle$ element in the $M \times M$ co-occurrence matrix. Table 1 lists the top-15 concepts that are statistically-related to `keratoconus` using the co-occurrence measure. Here, the co-occurrence measure is computed from the OHSUMED corpus which will be described in detail in Section 4.1.

3.2 Deriving Scenario-Specific Expansion Concepts

Using a statistical expansion method, we can derive a set of concepts that are statistically-related to the key concept, c_{key}, of the given query. Only a subset of these concepts are relevant to the given query's scenario, e.g., `treatment`. For example, the 5th and 8th concepts in Table 1, which are `acute hydrops` and `corneal`, are not related to the treatment of `keratoconus`. Therefore, in terms of deriving expansion concepts for query `keratoconus treatment`, these concepts should be filtered out. In this section, we will first describe the type of knowledge structure that enables us to perform this filtering and then present the filtering procedure.

UMLS - The Knowledge Source. *Unified Medical Language System* (*UMLS*) is a standard medical knowledge source developed by the National Library of Medicine. It consists of the *Metathesaurus*, the *Semantic Network*, and the *SPECIALIST lexicon*. The Semantic Network provides the essential knowledge structures for deriving scenario-specific expansion concepts, and is the primary focus of the following discussion. The Metathesaurus, which defines over 800K medical concepts and the hypernym/hyponym relationships among them, is used in our study for two purposes: 1) detecting concepts in both queries and document and 2) expanding hypernyms/hyponyms of a query's key concept. The second purpose will be further illustrated in Section 3.3. The lexicon is mainly used for unifying lexical features in medical-text-related natural language processing (NLP) and is not used in our study.

The Semantic Network defines about one hundred *semantic types* such as `Disease or Syndrome`, `Body Part`, etc. Each semantic type corresponds to a class/category of concepts. The semantic type of `Disease or Syndrome`, for instance, corresponds to 44,000 concepts in the Metathesaurus such as `keratoconus`, `lung cancer`, `diabetes`, etc. Besides the list of semantic types, the Semantic Network also defines the relations among various semantic types, such as `treats` and `diagnoses`. Such relations link isolated semantic types into a graph/network structure. The top half of Figure 2 presents a fragment of this network, which includes all semantic types that have a `treats` relation with the semantic type `Disease or Syndrome`. Relations such as `treats` in Figure 2 should be interpreted as follows: Any concepts that belong to semantic type `Therapeutic or Preventive Procedure`, e.g., `penetrating keratoplasty` or `chemotherapy`, have the potential to `treat` concepts that belong to the semantic type `Disease or Syndrome`, e.g., `keratoconus` or `lung cancer`. However, it is not indicated whether such relations concretely exist between two concepts, e.g., a `treats` relation between `penetrating keratoplasty` and `lung cancer`.

A Knowledge-Based Method to Derive Scenario-Specific Expansion Concepts. Given the knowledge structure in the Semantic Network, the basic idea in identifying scenario-specific expansion concept is to use this knowledge structure to filter out statistically-correlated concepts which do not belong to the "desirable" semantic types. Let us illustrate this idea through Figure 2, using the `treatment` scenario as an example: In this figure, we start with the set of concepts that are statistically-related to `keratoconus`. Our goal in applying the knowledge structure is to identify that: 1) concepts such as `penetrating keratoplasty`, `contact lens` and `griffonia` have the scenario-specific relation, i.e., `treats`, with `keratoconus` and should be kept during expansion; 2) concepts such as `acute hydrops` and

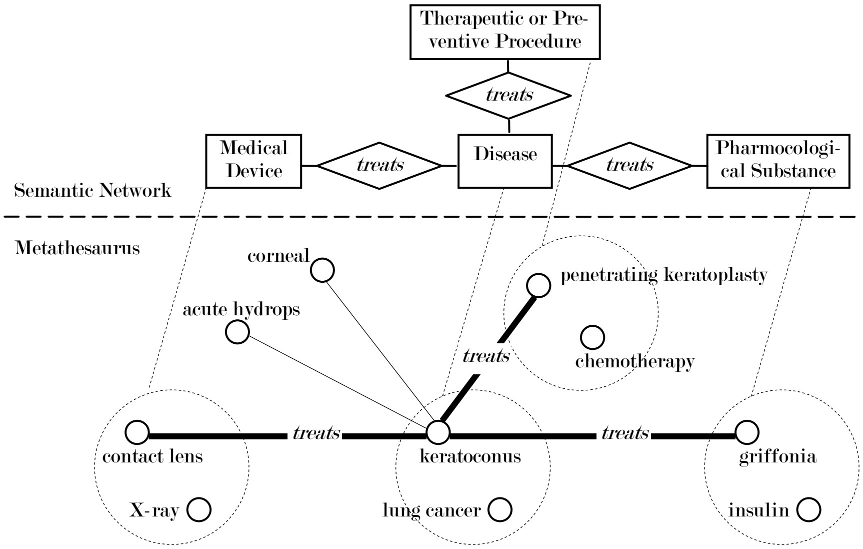


Figure 2: Using knowledge to identify scenario-specific concept relationships

`corneal` do not have the scenario-specific relation with `keratoconus` and should be filtered out.

In this figure, each solid circle represents one concept, and the solid lines connecting these solid circles indicate strong statistical correlations computed for a pair of concepts, e.g., the solid line between `keratoconus` and `contact lens`. A dotted circle represents a class of concepts, and a dotted line links that class of concepts to a corresponding semantic type. For example, concepts `keratoconus` and `lung cancer` are in the class that links to `Disease or Syndrome`.

We identified scenario-specific expansion concepts using the following process: Given a key concept c_{key} of the given query, we first identified the semantic type that c_{key} belongs to. For example, we identified `Disease or Syndrome` given the key concept `keratoconus`. Starting from that semantic type, we further followed the relations marked by the query's scenario and reached a set of relevant semantic types. For the previous example, given the query's scenario, `treatment`, we followed the `treats` relation to reach the three other semantic types as shown in Figure 2. Finally, we identified those statistically-related concepts that belong to the relevant semantic types as scenario specific. We further filtered out other statistically-related concepts which do not satisfy this criteria. From the previous example, this final step identified `penetrating keratoplasty`, `contact lens` and `griffonia` as scenario-specific expansion concepts and filtered out non-scenario-specific ones such as `acute hydrops` and `corneal`.

Table 2(a) and Table 2(b) show the lists of the concepts that `treat` and `diagnose keratoconus`, respectively. We derived these concepts based on the process

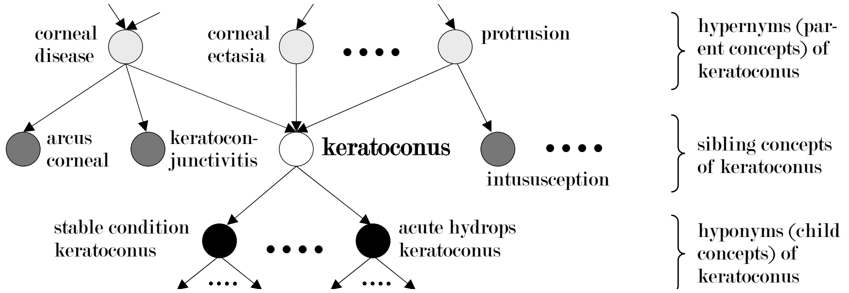

corneal disease
corneal ectasia
protrusion
hypernyms (parent concepts) of keratoconus
arcus corneal
keratoconjunctivitis
keratoconus
intussusception
sibling concepts of keratoconus
stable condition keratoconus
acute hydrops keratoconus
hyponyms (child concepts) of keratoconus

Figure 3: The direct parents, direct children and siblings for `keratoconus`

described above and show the top-15 concepts in terms of their correlation with `keratoconus`. To highlight the effectiveness in applying the knowledge-based filtering process, we can compare the concepts in Table 2 with those in Table 1 that are statistically correlated with `keratoconus`. 5 out of these 15 statistically-correlated concepts are kept in Table 2(a), whereas 2 are kept in Table 2(b). This comparison suggests that the knowledge structure is effective in filtering out concepts that are not closely related to the scenarios of `treatment` or `diagnosis`.

3.3 Hypernym/Hyponym Expansion

The goal of knowledge-based query expansion is to append specialized terms that appear in relevant documents but not in the original query. Scenario-specific concepts derived from the previous subsection represent a subset of such specialized terms. Another set of highly relevant terms contains hypernym/hyponyms of the key concept c_{key}.[6] For example, `corneal estasia`, a hypernym of `keratoconus`, is frequently mentioned by documents regarding `keratoconus, treatment`. Therefore, our technique should also expand those concepts that are close to c_{key} in the hypernym/hyponym hierarchy.

To expand hypernyms/hyponyms of the key concept to the original query, we again refer to the UMLS knowledge source. The Metathesaurus component defines not only the concepts but also the hypernym/hyponym relationships among these concepts. For example, Figure 3 shows the hypernyms (parents), hyponyms (children) and siblings of concept `keratoconus`, where the siblings of a concept is defined as those concepts that share the same parents with the given concept. Through empirical study (which will be discussed later), we have found that expanding the direct parents, direct children and siblings to the original query generates the best retrieval performance. This is in comparison to expanding parents/children that are two or more levels away from the key concept. Therefore, in the rest of our discussion, we will focus on expanding only the direct parents/children and siblings.

[6] A hypernym of concept c is a concept with a broader meaning than c, whereas a hyponym is one with a narrower meaning.

Concepts that statistically correlate to `keratoconus`	Weight
fuchs dystrophy	**0.289**
penetrating keratoplasty	**0.247**
epikeratoplasty	**0.230**
corneal ectasia	**0.168**
acute hydrops	**0.165**
keratometry	**0.133**
corneal topography	**0.132**
corneal	**0.130**
aphakic corneal edema	**0.122**
epikeratophakia	**0.119**
granular dystrophy corneal	**0.109**
keratoplasty	**0.103**
central cornea	**0.103**
contact lens	**0.101**
ghost vessels	**0.095**

(a)

Concepts that treat `keratoconus`	Weight
penetrating keratoplasty	**0.247**
epikeratoplasty	**0.230**
epikeratophakia	**0.119**
keratoplasty	**0.103**
contact lens	**0.101**
thermokeratoplasty	**0.092**
button	**0.067**
secondary lens implant	**0.057**
fittings adapters	**0.048**
esthesiometer	**0.043**
griffonia	**0.035**
trephine	**0.033**
slit lamps	**0.032**
fistulization	**0.030**
soft contact lens	**0.026**

(b)

Table 3: Weights for sample expansion concepts

3.4 Weight Adjustment for Expansion Terms

To match a query and a document using the Vector Space Model (VSM), we represent both the query and the document as vectors. Each term in the query becomes a dimension in the query vector, and receives a weight that quantifies the importance of this term in the entire query. Under this model, any additional term appended to the original query needs to be assigned a weight. An appropriate weight scheme for these additional terms is important because “under-weighting” will make the additional terms insignificant compared to the original query and lead to minor changes in the ranking of the retrieval results. On the contrary, “over-weighting” will make the additional terms improperly significant and cause a “topic drift” for the original query.

In the past, researchers have proposed weighting schemes for these additional terms based on the following intuition: The weight for an additional term c_a should be proportional to its correlation with the original query terms. In our problem the weight for c_a, w_a, is proportional to its correlation with the key concept c_{key}, i.e.:

$$w_a = co(c_a, c_{key}) \cdot w_{key} \tag{1}$$

In Eq.(1), the correlation between c_a and c_{key}, $co(c_a, c_{key})$, is derived using methods described in Section 3.1. w_{key} denotes the weight assigned to the key concept c_{key}. In Section 4.1 we will further explain how w_{key} is decided according to a common weighting scheme. Given that $co(c_a, c_{key})$ lies in $[0, 1]$, the weight that c_a receives will not exceed that of c_{key}. Using this equation, we compute the weights for the terms that statistically correlate with `keratoconus` (Table 1) and the weights for those that `treat keratoconus` (Table 2(a)). We list the weights for these terms in Table 3(a) and Table 3(b), respectively. These weights are computed by assuming the weight of the key concept (i.e., w_{key}) `keratoconus` is 1.

Weight Boosting. In our experiments we will compare the retrieval effectiveness of knowledge-based query expansion with that of statistical expansion. Since the knowledge-based method applies a filtering step to derive a subset of all statistically-related terms, the impact created by this subset on retrieval effectiveness will be less than the entire set of statistically-related terms. Therefore, weight adjustments are needed to compensate for the filtering. For instance, in our example of `keratoconus, treatment`, the "cumulative weight" for all terms in Table 3(b) is obviously smaller than the "cumulative weight" of those in Table 3(a). To increase the impact of the terms derived by the knowledge-based method, we can "boost" their weights by multiplying a linear factor β, so that the cumulative weight of those terms is comparable to those of the statistical-related terms. We refer to β as the *boosting factor*. With this factor, we alter Eq.(1) which assigns the weight for any additional term c_a as follows:

$$w_a = \beta \cdot co(c_a, c_{key}) \cdot w_{key} \tag{2}$$

We derive β based on the following intuition. We quantify the cumulative weight for both the statistical expansion terms (e.g., those in Table 3(a)) and the knowledge-based expansion terms (e.g., those in Table 3(b)). The former cumulative weight will be larger than the latter. We define β to be the former divided by the latter. In this way, the cumulative weight for the knowledge-based expansion terms equals to that of the statistical expansion terms after boosting.

More specifically, we quantify the cumulative weight of a set of expansion terms using the length of the "expansion vector" composed by these terms. Here we define the vector length according to the standard vector space notation: Let $V^{KB} = \langle w_1^{KB}, ..., w_k^{KB} \rangle$ be the augmenting vector consisting solely of terms derived by the knowledge-based method, where $w_i^{KB} (1 \leq i \leq k)$ denotes the weight for the i_{th} term in knowledge-based expansion (Eq.(1)). Likewise, let $V^{stat} = \langle w_1^{stat}, ..., w_l^{stat} \rangle$ be the augmenting vector consisting of all statistically related terms. The process of deriving $\{w_1^{KB}, ..., w_k^{KB}\}$ yields $k < l$. Consequently, $\{w_1^{KB}, ..., w_k^{KB}\} \subset \{w_1^{stat}, ..., w_l^{stat}\}$. Let $|V^{KB}|$ be the length of the vector V^{KB}, i.e.,

$$|V^{KB}| = \sqrt{(w_1^{KB})^2 + (w_2^{KB})^2 + \cdots + (w_k^{KB})^2} \tag{3}$$

Likewise, let $|V^{stat}|$ represent the length of vector V^{stat} which can be computed similarly as Eq.(3). Further, we define the *boosting factor* for V^{KB} to be:

$$\beta = \frac{|V^{stat}|}{|V^{KB}|} \tag{4}$$

In our experiments, we will experimentally study the effects of boosting by comparing the retrieval results with and without using boosting. Furthermore, we are interested in studying the effects of different levels of boosting to gain insight on the "optimal" boosting level. This motivates us to introduce a *boosting-level-controlling factor* α to refine Eq.(4):

$$\beta_r = 1 + \alpha \cdot (\frac{|V^{stat}|}{|V^{KB}|} - 1) \tag{5}$$

where β_r is the refined boosting factor. The parameter α, ranging within $[0, 1]$, can be used to control the boosting scale. From Eq.(5), we note that $\beta_r = 1$ when we set $\alpha = 0$, which represents no boosting. β_r increases as α increases. As α increases to 1, β_r becomes $\frac{|V^{stat}|}{|V^{KB}|}$. (In our experiments, we have actually evaluated cases by setting $\alpha > 1$. As the results will show later, the retrieval effectiveness is usually suboptimal compared to an α value within $[0, 1]$.) Thus, we can use α to experimentally study the sensitivity of retrieval results with regard to boosting.

4 Experimental Results

In this section, we experimentally evaluate the effectiveness of the knowledge-based query expansion for two standard medical corpuses. Our main focus is to compare the results of our technique with that of statistical expansion. We start with the experiment setup and then present the results under selective settings.

4.1 Experiment Setup

4.1.1 Testbeds

A testbed for a retrieval experiment consists of three components: 1) a corpus (or a document collection), 2) a set of benchmark queries and 3) relevance judgments indicating which documents are relevant for each query. Our experiment is based on the following two testbeds:

OHSUMED [Hersh et al.(1994)]. This testbed 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 the abstracts of 348,000 MEDLINE articles from 1988 to 1992. Each document contains a title, the abstract, a set of Medical Subject Headings (MeSH), author information, publication type, source, a MEDLINE identifier, and a document ID. The MeSH headings are expert-assigned indexing terms drawn from a subset of UMLS concepts. In our experiment, we only keep the title and the abstract in representing each document. We discard the MeSH headings in order to simulate a typical information retrieval setting in which no expert-assigned indexing terms are available.
- Query set. The query set consists of 106 queries. Each query contains a patient description, an information request and a query ID. We are interested in short and general queries. Thus, we use the information-request sub-portion to represent each query. Among the 106 queries, we have identified a total number of 57 queries that are scenario-specific. In Table 4, we categorize these 57 queries based on the scenario(s) each query mentions. The corresponding ID of each query is listed in this table. (The full text of each query is shown

Scenario	Query IDs
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

Table 4: IDs of OHSUMED queries mentioning each scenario

in [Liu and Chu(2006)]). Note that a query mentioning multiple distinct scenarios will appear multiple times in this table corresponding to its scenarios.

- Relevance judgments. 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 [Hersh et al.(1994)]. In our experiments, we restrict the retrieval to the 14,430 judged documents only and count both the DR and the PR documents as relevant answers as we measure the precision-recall of a particular retrieval method.[7]

The McMaster Clinical HEDGES Database [Wilczynski et al.(2001), Wong et al.(2003), Wilczynski and Haynes(2003), Montori et al.(2003)]. This testbed was originally constructed for the task of medical document classification instead of free-text query answering. As a result, adaptation is needed for our study. We will first describe the original dataset, and then explain how we adapted it to make it a usable testbed for our experimental evaluation.

- Original dataset. The McMaster Clinical HEDGES Database contains 48,000 PubMed articles published in 2000. Each article was classified into the following scenario categories: *treatment*, *diagnosis*, *etiology*, *prognosis*, *clinical prediction guide* of a disease, *economics* of a healthcare issue, or *review* of a healthcare topic. Consensus about the classification was drawn among six human experts [Wilczynski et al.(2001)]. When the experts classified each article, they had access to the hardcopies of the full text. However, to construct a testbed for our retrieval system, we were only able to download the title and abstract of

[7] Treating both the DR and the PR documents as relevant documents is consistent with the settings in existing studies [Hersh et al.(1994), Hersh et al.(2000)]

each article from the PubMed system. (The full text of each article is typically unavailable through PubMed.)

- Construction of Scenario-Specific Queries. Since the McMaster Clinical HEDGES Database is constructed to test document classification, it does not contain a query set. Using the following procedure, we constructed a set of 55 scenario-specific queries, and determined the relevance judgments for these queries based on the document classification that can be adapted for these queries:

  Step 1. We identified all the disease/symptom concepts in the OHSUMED query set. We identified such concepts based on their semantic type information (defined by UMLS). We used these concepts as the key concepts in constructing our scenario-specific queries for the McMaster testbed. In selecting these concepts, we manually filtered out eight concepts (out of an original number of 90 concepts) that we considered as too general to make a scenario-specific query, e.g., `infection`, `lesion` and `carcinoma`. After this step, we obtained 82 such key concepts.

  Step 2. For each key concept identified in Step 1, we constructed four scenario-specific queries, namely the `treatment`, `diagnosis`, `etiology` and `prognosis` of a disease/symptom. For example, for the concept `breast cancer`, we constructed the queries `breast cancer treatment`, `breast cancer diagnosis`, `breast cancer etiology`, and `breast cancer prognosis`. We restricted our study to these four scenarios because our current knowledge source only covers these four scenarios.

  Step 3. For each query generated in Step 2, we generated its relevance judgments by applying the following simple criterion: A document is considered to be relevant to a given query if 1) experts have classified the document to the category of the query's scenario and 2) the document mentions the query's key concept. This criterion has been our best choice to automate the process of generating relevance judgments on a relatively large scale; however, it may misidentify irrelevant documents as relevant. After we identified the relevant documents for each query, we further filtered out certain queries based on the intuition that a query with too few relevance judgments will lead to less reliable retrieval results (especially in terms of precision/recall). For example, for a query with only one relevant document, two similar retrieval systems may obtain completely different precision/recall results if one ranks the relevant document on top, and another accidentally ranks it out of top-10. To implement this intuition, we filtered out queries that have less than 5 relevant documents. After this filtering step, we were left with 55 queries.

  Due to space limitations, we show the 55 McMaster queries together with the scenarios identified for each query in the extended version of this paper [Liu and Chu(2006)].

4.1.2 VSM and Indexing

In Information Retrieval studies, *indexing* typically refers to the step of converting free-text documents and queries to representations comprehensible to a query-document similarity computation model, e.g., the Vector Space Model (VSM) [Salton and McGill(1983)] or a probabilistic retrieval model [Callan et al.(1992)]. In our study, we focused on experimental evaluation using the stem-based VSM [Salton and McGill(1983)], a VSM that is extensively applied in similar studies.

Using a stem-based VSM, both a query and a document are represented as vectors of word stems. Given a piece of free text, we first removed common stop words such as "a," "the," etc., and then derived word stems from the text using the Lovins stemmer [Lovins(1968)]. We further applied the $tf \cdot idf$ weighting scheme (more specifically the $atc \cdot atc$ scheme [Salton and Buckley(1988)]) to assign weights to stems in documents and the query before expansion. (This weighting process yields the weight for the key concept in Eq.(1).

Under the stem-based VSM, all terms expanded to a given query need to be in the word-stem format. Thus, for expansion concepts derived from procedures in Section 3.2 and Section 3.3, we applied the following procedure to identify the corresponding word stems: For each expansion concept, we first looked up its string forms in UMLS. We further removed stop words and used the Lovins stemmer to convert the string forms into word stems. Lastly, we assigned weights to these expansion word stems using the method described in Section 3.4.

4.2 Retrieval Performance

In the following, we study the performance improvement of knowledge-based expansion compared to that of statistical expansion. We first study the improvements for selected expansion sizes, then study the sensitivity of boosting for selected query scenarios.

The retrieval performance is measured using the following three different metrics:

avgp - 11-point precision average (precision averaged over the 11 standard recall points [Salton and McGill(1983)])

p@10 - precision in top-10 retrieved documents

p@20 - precision in top-20 retrieved documents

Expansion Sizes. For a given expansion size s, we used both knowledge-based expansion and statistical expansion to expand the top-s stems that have the heaviest weights. For knowledge-based expansion, no weight boosting was applied at this stage.

We compute the three metrics for both methods on the OHSUMED and McMaster testbeds. We further average the results over the queries in these two testbeds. Table 5

s	10	20	30	40	50	100	200	300	All
Statistical Expansion	0.417	0.424	0.428	0.43	0.429	0.432	0.429	0.43	0.425
Knowledge-Based Expansion (% of improvement)	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%)

(a) Performance comparison using the *avgp* metric for the OHSUMED testbed

s	10	20	30	40	50	100	200	300	All
Statistical Expansion	0.535	0.546	0.549	0.553	0.551	0.567	0.581	0.574	0.567
Knowledge-Based Expansion (% of improvement)	0.544 (1.7%)	0.547 (0.2%)	0.554 (1.0%)	0.551 (-0.4%)	0.553 (0.4%)	0.572 (0.9%)	0.572 (-1.5%)	0.577 (0.5%)	0.588 (3.7%)

(b) Performance comparison using the *p@10* metric for the OHSUMED testbed

s	10	20	30	40	50	100	200	300	All
Statistical Expansion	0.482	0.491	0.493	0.491	0.492	0.496	0.497	0.493	0.496
Knowledge-Based Expansion (% of improvement)	0.483 (0.2%)	0.491 (0%)	0.494 (0.2%)	0.496 (1%)	0.493 (0.2%)	0.498 (0.4%)	0.496 (-0.2%)	0.497 (0.8%)	0.498 (0.4%)

(c) Performance comparison using the *p@20* metric for the OHSUMED testbed

s	10	20	30	40	50	100	200	300	All
Statistical Expansion	0.326	0.328	0.325	0.324	0.323	0.319	0.311	0.309	0.295
Knowledge-Based Expansion (% of improvement)	0.325 (-0.1%)	0.328 (0.1%)	0.324 (-0.3%)	0.326 (0.8%)	0.325 (0.4%)	0.324 (1.4%)	0.321 (3.3%)	0.32 (3.4%)	0.321 (9%)

(d) Performance comparison using the *avgp* metric for the McMaster testbed

s	10	20	30	40	50	100	200	300	All
Statistical Expansion	0.316	0.324	0.324	0.318	0.324	0.311	0.295	0.3	0.293
Knowledge-Based Expansion (% of improvement)	0.322 (1.7%)	0.324 (0%)	0.322 (-0.6%)	0.325 (2.3%)	0.322 (-0.6%)	0.318 (2.3%)	0.315 (6.8%)	0.32 (6.7%)	0.335 (14.3%)

(e) Performance comparison using the *p@10* metric for the McMaster testbed

s	10	20	30	40	50	100	200	300	All
Statistical Expansion	0.285	0.285	0.285	0.283	0.283	0.281	0.279	0.278	0.279
Knowledge-Based Expansion (% of improvement)	0.285 (0.3%)	0.287 (0.6%)	0.287 (1%)	0.291 (2.9%)	0.29 (2.6%)	0.293 (4.2%)	0.286 (2.6%)	0.291 (4.6%)	0.292 (4.6%)

(f) Performance comparison using the *p@20* metric for the McMaster testbed

Table 5: Performance comparison of the two expansion methods under various expansion sizes

shows the performance comparison of the two methods on both testbeds, under various metrics. The first row in each subtable shows the performance of statistical expansion, whereas the second row shows that of knowledge-based expansion and its percentage of improvement over statistical expansion.

In these figures, "s=All" means appending all possible expansion terms that have a non-zero weight (Eq.(2)) into the original query. Using the knowledge-based method, setting "s=All" led to expanding 1717 terms to each query on average, with standard deviation as 1755; using the statistical method, it led to an average of 50317 terms and 15243 being the standard deviation.

From these experimental results, we observe the following: The performance for knowledge-based expansion generally increases as s increases and usually reaches the peak when s=All. (The only exception is in the case of using the *avgp* metric on the McMaster testbed, in which the performance of the knowledge-based method roughly remains stable as s increases.) On the other hand, the performance of the statistical method degrades as s becomes larger. On the OHSUMED testbed, its performance de-

α \ s	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%)

Table 6: Weight boosting for the OHSUMED testbed, measured by *avgp*

grades after s=100 (Table 5(a)) or s=200 (Table 5(b) and Table 5(c)); on the McMaster testbed, the performance starts degrading almost immediately after s becomes greater than 20. This is due to the fact that statistical expansion does not distinguish between expansion terms that are scenario-specific and those that are not. As a result, as more terms are appended to the original query, the negative effect of including 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.

We have also compared the statistical expansion method with no expansion, to understand the general effectiveness of query expansion on the scenario-specific queries we chose. Due to space limit, we have not included the result of this comparison. In general, statistical expansion consistently outperforms the no-expansion method by more than 5%, which represents a significant improvement. In other words, the method of statistical expansion that we are comparing against already generates reasonably good retrieval results.)

The Effectiveness of Weight Boosting. In the next experiments, we multiplied a boosting factor to the weights of knowledge-based expansion terms (Eq.(2)). The boosting factor β is computed using Eq.(5), under the different settings of $\alpha = 0.25, 0.5, 0.75, 1, 1.25, 1.5$. Table 6 to Table 11 show the effects of different boosting amounts on the performance for knowledge-based query expansion, under the three metrics and for the two testbeds. Each cell in these tables shows 1) the performance of knowledge-based expansion and 2) the percentage of improvement of knowledge-based expansion over statistical expansion under the same expansion size. In these tables, the thick-bordered cells represent the best performance for that column (i.e. under the same setting of expansion size); shaded cells represent the best performance for that row (i.e. under the same setting of boosting factor). The best performance in the entire table is highlighted in the shaded and thick-bordered cell.

α \ s	10	20	30	40	50	100	200	300	All
0 (no boosting)	0.544 (1.7%)	0.547 (0.2%)	0.554 (1.0%)	0.551 (-0.4%)	0.553 (0.4%)	0.572 (0.9%)	0.572 (-1.5%)	0.577 (0.5%)	0.588 (3.7%)
0.25	0.549 (2.6%)	0.556 (1.8%)	0.556 (1.3%)	0.558 (0.9%)	0.567 (2.9%)	0.572 (0.9%)	0.577 (-0.7%)	0.588 (2.4%)	0.595 (4.9%)
0.5	0.549 (2.6%)	0.561 (2.7%)	0.563 (2.6%)	0.565 (2.2%)	0.570 (3.4%)	0.584 (3%)	0.586 (0.9%)	0.593 (3.3%)	0.6 (5.8%)
0.75	0.546 (2.1%)	0.565 (3.5%)	0.561 (2.2%)	0.568 (2.7%)	0.568 (3.1%)	0.589 (3.9%)	0.584 (0.5%)	0.595 (3.7%)	0.596 (5.1%)
1	0.552 (3.2%)	0.567 (3.8%)	0.568 (3.5%)	0.577 (4.3%)	0.577 (4.7%)	0.595 (4.9%)	0.586 (0.9%)	0.595 (3.7%)	0.6 (5.8%)
1.25	0.554 (3.6%)	0.560 (2.6%)	0.567 (3.3%)	0.567 (2.5%)	0.572 (3.8%)	0.582 (2.6%)	0.579 (-0.3%)	0.591 (3%)	0.593 (4.6%)
1.5	0.558 (4.3%)	0.558 (2.2%)	0.570 (3.8%)	0.570 (3.1%)	0.574 (4.2%)	0.581 (2.5%)	0.577 (-0.7%)	0.588 (2.4%)	0.584 (3%)

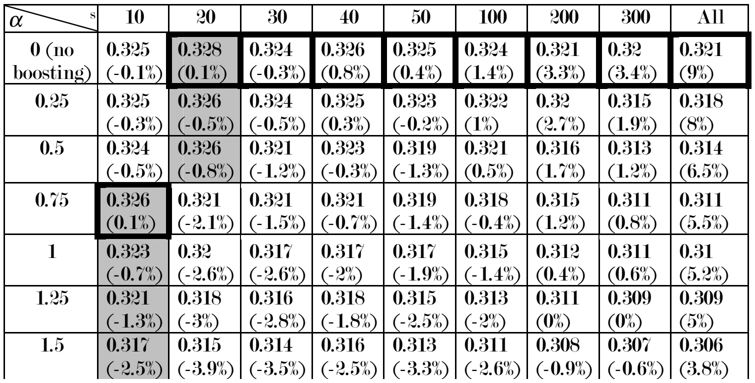
Table 7: Weight boosting for the OHSUMED testbed, measured by *p@10*

α \ s	10	20	30	40	50	100	200	300	All
0 (no boosting)	0.483 (0.2%)	0.491 (0%)	0.494 (0.2%)	0.496 (1%)	0.493 (0.2%)	0.498 (0.4%)	0.496 (-0.2%)	0.497 (0.8%)	0.498 (0.4%)
0.25	0.486 (0.8%)	0.496 (1%)	0.494 (0.2%)	0.499 (1.6%)	0.496 (0.8%)	0.503 (1.4%)	0.502 (1%)	0.503 (2%)	0.502 (1.2%)
0.5	0.486 (0.8%)	0.499 (1.6%)	0.499 (1.2%)	0.503 (2.4%)	0.502 (2%)	0.509 (2.6%)	0.509 (2.4%)	0.511 (3.7%)	0.511 (3%)
0.75	0.487 (1%)	0.496 (1%)	0.499 (1.2%)	0.509 (3.7%)	0.507 (3%)	0.510 (2.8%)	0.512 (3%)	0.510 (3.4%)	0.511 (3%)
1	0.483 (0.2%)	0.498 (1.4%)	0.501 (1.6%)	0.509 (3.7%)	0.507 (3%)	0.511 (3%)	0.517 (4%)	0.512 (3.9%)	0.510 (2.8%)
1.25	0.482 (0%)	0.496 (1%)	0.498 (1%)	0.510 (3.9%)	0.509 (3.5%)	0.514 (3.6%)	0.514 (3.4%)	0.513 (4.1%)	0.511 (3%)
1.5	0.487 (1%)	0.492 (0.2%)	0.498 (1%)	0.508 (3.5%)	0.504 (3.4%)	0.513 (3.4%)	0.513 (3.2%)	0.511 (3.7%)	0.507 (2.2%)

Table 8: Weight boosting for the OHSUMED testbed, measured by *p@20*

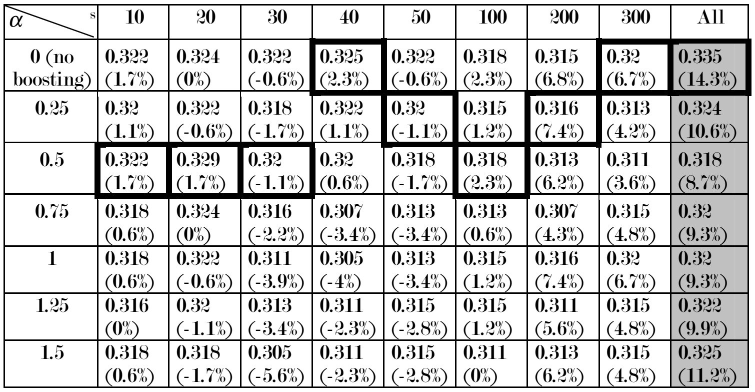
The following observations can be made from these results:

- For the OHSUMED testbed, the best performance within each column (the thick-bordered cells) generally falls in the range from $\alpha = 0.5$ to $\alpha = 1.25$. This indicates that boosting helps improve the performance of knowledge-based expansion. In particular, boosting introduces significant improvements under the metrics of *p@10* and *p@20*. We note that setting $\alpha = 0.5$ or $= 0.75$ generally yields the best boosting effect for the *avgp* metric; setting $\alpha = 1$ or $= 1.25$ yields better performances for the *p@10* and *p@20* metrics.
- For the McMaster testbed, however, boosting seems to be less effective: The best performance within each column falls in the range from $\alpha = 0$ to $\alpha = 0.75$.
- For both testbeds, if we fix the boosting factor, the best performance within each row (the shaded cells) is generally achieved by having an expansion size s as large as possible (with the exception case of the *avgp* metric measured on McMaster). This is consistent with the reported results in the previous experiments.



α s	10	20	30	40	50	100	200	300	All
0 (no boosting)	0.325 (-0.1%)	0.328 (0.1%)	0.324 (-0.3%)	0.326 (0.8%)	0.325 (0.4%)	0.324 (1.4%)	0.321 (3.3%)	0.32 (3.4%)	0.321 (9%)
0.25	0.325 (-0.3%)	0.326 (-0.5%)	0.324 (-0.5%)	0.325 (0.3%)	0.323 (-0.2%)	0.322 (1%)	0.32 (2.7%)	0.315 (1.9%)	0.318 (8%)
0.5	0.324 (-0.5%)	0.326 (-0.8%)	0.321 (-1.2%)	0.323 (-0.3%)	0.319 (-1.3%)	0.321 (0.5%)	0.316 (1.7%)	0.313 (1.2%)	0.314 (6.5%)
0.75	0.326 (0.1%)	0.321 (-2.1%)	0.321 (-1.5%)	0.321 (-0.7%)	0.319 (-1.4%)	0.318 (-0.4%)	0.315 (1.2%)	0.311 (0.8%)	0.311 (5.5%)
1	0.323 (-0.7%)	0.32 (-2.6%)	0.317 (-2.6%)	0.317 (-2%)	0.317 (-1.9%)	0.315 (-1.4%)	0.312 (0.4%)	0.311 (0.6%)	0.31 (5.2%)
1.25	0.321 (-1.3%)	0.318 (-3%)	0.316 (-2.8%)	0.318 (-1.8%)	0.315 (-2.5%)	0.313 (-2%)	0.311 (0%)	0.309 (0%)	0.309 (5%)
1.5	0.317 (-2.5%)	0.315 (-3.9%)	0.314 (-3.5%)	0.316 (-2.5%)	0.313 (-3.3%)	0.311 (-2.6%)	0.308 (-0.9%)	0.307 (-0.6%)	0.306 (3.8%)

Table 9: Weight boosting for the McMaster testbed, measured by *avgp*

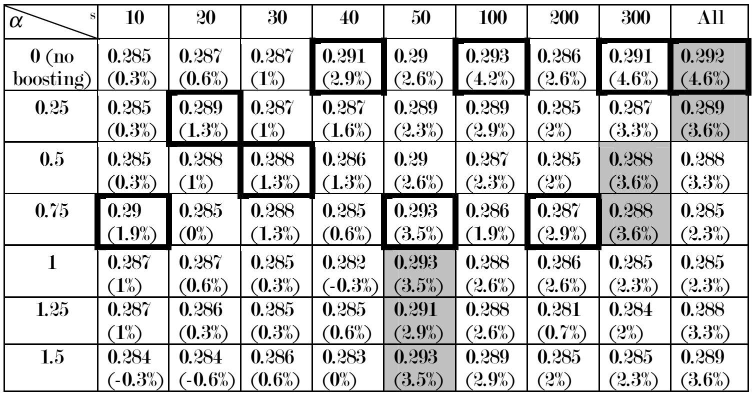


α s	10	20	30	40	50	100	200	300	All
0 (no boosting)	0.322 (1.7%)	0.324 (0%)	0.322 (-0.6%)	0.325 (2.3%)	0.322 (-0.6%)	0.318 (2.3%)	0.315 (6.8%)	0.32 (6.7%)	0.335 (14.3%)
0.25	0.32 (1.1%)	0.322 (-0.6%)	0.318 (-1.7%)	0.322 (1.1%)	0.32 (-1.1%)	0.315 (1.2%)	0.316 (7.4%)	0.313 (4.2%)	0.324 (10.6%)
0.5	0.322 (1.7%)	0.329 (1.7%)	0.32 (-1.1%)	0.32 (0.6%)	0.318 (-1.7%)	0.318 (2.3%)	0.313 (6.2%)	0.311 (3.6%)	0.318 (8.7%)
0.75	0.318 (0.6%)	0.324 (0%)	0.316 (-2.2%)	0.307 (-3.4%)	0.313 (-3.4%)	0.313 (0.6%)	0.307 (4.3%)	0.315 (4.8%)	0.32 (9.3%)
1	0.318 (0.6%)	0.322 (-0.6%)	0.311 (-3.9%)	0.305 (-4%)	0.313 (-3.4%)	0.315 (1.2%)	0.316 (7.4%)	0.32 (6.7%)	0.32 (9.3%)
1.25	0.316 (0%)	0.32 (-1.1%)	0.313 (-3.4%)	0.311 (-2.3%)	0.315 (-2.8%)	0.315 (1.2%)	0.311 (5.6%)	0.315 (4.8%)	0.322 (9.9%)
1.5	0.318 (0.6%)	0.318 (-1.7%)	0.305 (-5.6%)	0.311 (-2.3%)	0.315 (-2.8%)	0.311 (0%)	0.313 (6.2%)	0.315 (4.8%)	0.325 (11.2%)

Table 10: Weight boosting for the McMaster testbed, measured by *p@10*

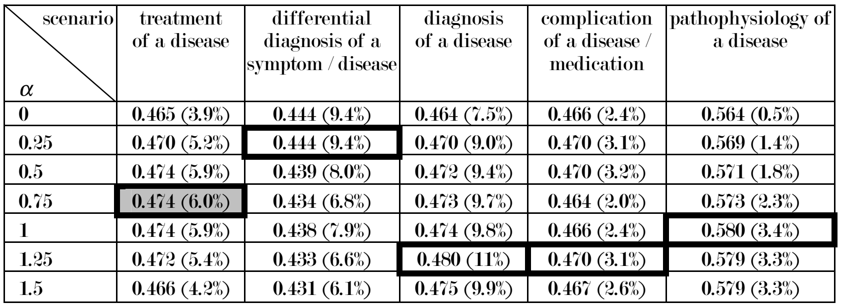
Sensitivity of Performance Improvements with Query Scenarios. We now study how knowledge-based expansion perform for different query scenarios. For the OHSUMED testbed, we grouped the 57 queries according to their scenarios, and further selected the five largest groups of scenarios, namely `treatment`, `diagnosis`, `pathophysiology` of a disease, `differential diagnosis` of a symptom/disease and `complications` of a disease/medication. We skipped the remaining scenarios because each of these scenarios has too few queries to derive reliable statistics. (The number of queries that belong to each scenario can be easily counted from Table 4.) Similarly, we grouped the 55 McMaster queries based on the four scenarios they belong to: namely `treatment`, `diagnosis`, `etiology` and `prognosis` of a disease.

We average the performance of knowledge-based expansion within each group of queries and show the *avgp*, *p@10* and *p@20* results in Table 12 to Table 14. Each cell in these tables shows 1) the performance of knowledge-based expansion averaged



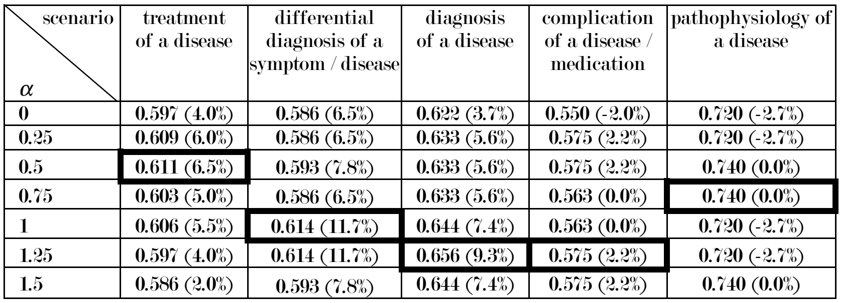
α s	10	20	30	40	50	100	200	300	All
0 (no boosting)	0.285 (0.3%)	0.287 (0.6%)	0.287 (1%)	0.291 (2.9%)	0.29 (2.6%)	0.293 (4.2%)	0.286 (2.6%)	0.291 (4.6%)	0.292 (4.6%)
0.25	0.285 (0.3%)	0.289 (1.3%)	0.287 (1%)	0.287 (1.6%)	0.289 (2.3%)	0.289 (2.9%)	0.285 (2%)	0.287 (3.3%)	0.289 (3.6%)
0.5	0.285 (0.3%)	0.288 (1%)	0.288 (1.3%)	0.286 (1.3%)	0.29 (2.6%)	0.287 (2.3%)	0.285 (2%)	0.288 (3.6%)	0.288 (3.3%)
0.75	0.29 (1.9%)	0.285 (0%)	0.288 (1.3%)	0.285 (0.6%)	0.293 (3.5%)	0.286 (1.9%)	0.287 (2.9%)	0.288 (3.6%)	0.285 (2.3%)
1	0.287 (1%)	0.287 (0.6%)	0.285 (0.3%)	0.282 (-0.3%)	0.293 (3.5%)	0.288 (2.6%)	0.286 (2.6%)	0.285 (2.3%)	0.285 (2.3%)
1.25	0.287 (1%)	0.286 (0.3%)	0.285 (0.3%)	0.285 (0.6%)	0.291 (2.9%)	0.288 (2.6%)	0.281 (0.7%)	0.284 (2%)	0.288 (3.3%)
1.5	0.284 (-0.3%)	0.284 (-0.6%)	0.286 (0.6%)	0.283 (0%)	0.293 (3.5%)	0.289 (2.9%)	0.285 (2%)	0.285 (2.3%)	0.289 (3.6%)

Table 11: Weight boosting for the McMaster testbed, measured by *p@20*



scenario α	treatment of a disease	differential diagnosis of a symptom / disease	diagnosis of a disease	complication of a disease / medication	pathophysiology of a disease
0	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%)

Table 12: Performance improvements for selected scenarios measured by *avgp* for the OHSUMED testbed. Expansion size s=All



scenario α	treatment of a disease	differential diagnosis of a symptom / disease	diagnosis of a disease	complication of a disease / medication	pathophysiology of a disease
0	0.597 (4.0%)	0.586 (6.5%)	0.622 (3.7%)	0.550 (-2.0%)	0.720 (-2.7%)
0.25	0.609 (6.0%)	0.586 (6.5%)	0.633 (5.6%)	0.575 (2.2%)	0.720 (-2.7%)
0.5	0.611 (6.5%)	0.593 (7.8%)	0.633 (5.6%)	0.575 (2.2%)	0.740 (0.0%)
0.75	0.603 (5.0%)	0.586 (6.5%)	0.633 (5.6%)	0.563 (0.0%)	0.740 (0.0%)
1	0.606 (5.5%)	0.614 (11.7%)	0.644 (7.4%)	0.563 (0.0%)	0.720 (-2.7%)
1.25	0.597 (4.0%)	0.614 (11.7%)	0.656 (9.3%)	0.575 (2.2%)	0.720 (-2.7%)
1.5	0.586 (2.0%)	0.593 (7.8%)	0.644 (7.4%)	0.575 (2.2%)	0.740 (0.0%)

Table 13: Performance improvements for selected scenarios measured by *p@10* for the OHSUMED testbed. Expansion size s=All

scenario / α	treatment of a disease	differential diagnosis of a symptom / disease	diagnosis of a disease	complication of a disease / medication	pathophysiology of a disease
0	0.497 (-0.3%)	0.500 (1.4%)	0.517 (-1.1%)	0.525 (-2.3%)	0.720 (0.0%)
0.25	0.506 (1.4%)	0.500 (1.4%)	0.539 (3.2%)	0.531 (-1.2%)	0.710 (-1.4%)
0.5	0.514 (3.2%)	0.500 (1.4%)	0.539 (3.2%)	0.538 (0.0%)	0.710 (-1.4%)
0.75	0.520 (4.3%)	0.500 (1.4%)	0.550 (5.3%)	0.538 (0.0%)	0.700 (-2.8%)
1	0.523 (4.9%)	0.507 (2.9%)	0.572 (9.6%)	0.544 (1.2%)	0.700 (-2.8%)
1.25	0.519 (4.0%)	0.507 (2.9%)	0.550 (5.3%)	0.544 (1.2%)	0.700 (-2.8%)
1.5	0.516 (3.4%)	0.507 (2.9%)	0.544 (4.3%)	0.544 (1.2%)	0.700 (-2.8%)

Table 14: Performance improvements for selected scenarios measured by *p@20* for the OHSUMED testbed. Expansion size s=200

scenario / α	treatment of a disease	diagnosis of a disease	etiology of a disease	prognosis of a disease
0	0.49 (-0.6%)	0.145 (3.9%)	0.324 (0.9%)	0.229 (-0.4%)
0.25	0.488 (-1%)	0.145 (3.8%)	0.319 (-0.7%)	0.23 (-0.1%)
0.5	0.486 (-1.4%)	0.143 (2.8%)	0.318 (-1%)	0.231 (0.1%)
0.75	0.484 (-1.9%)	0.143 (2.5%)	0.307 (-4.4%)	0.231 (0.2%)
1	0.48 (-2.7%)	0.142 (2.1%)	0.304 (-5.2%)	0.232 (0.6%)
1.25	0.477 (-3.3%)	0.142 (2%)	0.3 (-6.3%)	0.235 (1.8%)
1.5	0.47 (-4.7%)	0.142 (1.5%)	0.298 (-7.1%)	0.234 (1.7%)

Table 15: Performance improvements for selected scenarios measured by *avgp* for the McMaster testbed. Expansion size s=20

over the corresponding group of queries, under the corresponding boosting setting (α), and 2) the percentage of improvement of knowledge-based expansion over statistical expansion under the same settings. For example, the shaded cell in Table 12 shows that among the 35 `treatment` OHSUMED queries, 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 Table 12, Table 13 and Table 14, we set the expansion size s=All, All and 200, respectively; for the results in Table 15, Table 16 and Table 17, we set the expansion size s=20, All, and 50. Such settings are based on our observations in the previous subsection where the knowledge-based method tend to perform the best with these expansion sizes under the corresponding evaluation metrics.

These results generally suggest that knowledge-based expansion performs differently for queries with different scenarios. More specifically, the method yields more improvements in scenarios such as `treatment`, `differential diagnosis` and `diagnosis`, whereas it yields less improvements in such scenarios as `complication`, `pathophysiology`, `etiology` and `prognosis`. An explanation lies in the different knowledge structures for these scenarios. The knowledge

scenario / α	treatment of a disease	diagnosis of a disease	etiology of a disease	prognosis of a disease
0	0.482 (26.2%)	0.129 (12.5%)	0.324 (5.8%)	0.271 (5.6%)
0.25	0.465 (21.5%)	0.129 (12.5%)	0.318 (3.8%)	0.257 (0%)
0.5	0.447 (16.9%)	0.114 (0%)	0.324 (5.8%)	0.257 (0%)
0.75	0.441 (15.4%)	0.143 (25%)	0.318 (3.8%)	0.264 (2.8%)
1	0.441 (15.4%)	0.143 (25%)	0.318 (3.8%)	0.264 (2.8%)
1.25	0.435 (13.8%)	0.143 (25%)	0.318 (3.8%)	0.279 (8.3%)
1.5	0.441 (15.4%)	0.143 (25%)	0.318 (3.8%)	0.286 (11.1%)

Table 16: Performance improvements for selected scenarios measured by *p@10* for the McMaster testbed. Expansion size s=All

scenario / α	treatment of a disease	diagnosis of a disease	etiology of a disease	prognosis of a disease
0	0.462 (3.3%)	0.1 (7.7%)	0.282 (2.1%)	0.186 (0%)
0.25	0.462 (3.3%)	0.107 (15.4%)	0.276 (0%)	0.186 (0%)
0.5	0.462 (3.3%)	0.114 (23.1%)	0.276 (0%)	0.186 (0%)
0.75	0.468 (4.6%)	0.121 (30.8%)	0.271 (-2.1%)	0.193 (3.8%)
1	0.471 (5.3%)	0.114 (23.1%)	0.268 (-3.2%)	0.196 (5.8%)
1.25	0.462 (3.3%)	0.114 (23.1%)	0.268 (-3.2%)	0.2 (7.7%)
1.5	0.468 (4.6%)	0.114 (23.1%)	0.265 (-4.3%)	0.204 (9.6%)

Table 17: Performance improvements for selected scenarios measured by *p@20* for the McMaster testbed. Expansion size s=50

structures (i.e., the fragments of UMLS Semantic Network such as Figure 2) for the latter four scenarios were originally missing in UMLS and were acquired by ourselves from experts. (We will further present the details of this knowledge acquisition process in Section 5.) These acquired structures have more semantic types marked as relevant than those for the former three scenarios. As a result, when handling queries with the latter four scenarios, the knowledge-based method keeps more concepts during the filtering step. Thus, the expansion result for the knowledge-based method resembles that of the statistical expansion method, leading to almost equivalent performance between the two methods and less improvements. We believe that a refined clustering and ranking of the knowledge structures for the four scenarios (i.e., `complication`, `pathophysiology`, `etiology` and `prognosis`) will increase the improvements in retrieval performance.

4.3 Discussion of Results

Choice of α for weight boosting. Our experimental results from Table 6 to Table 11 suggest that weight boosting is helpful in improving retrieval performance. Further, the results shown in Table 12 to Table 17 suggest that the performance of weight boosting is sensitive to the query scenario. Certain query scenarios such as `treatment` and diagnosis are associated with more compact knowledge structures, which leads to significantly less expansion concepts using our knowledge-based method compared to those by statistical expansion. In these scenarios, setting α in between 0.75 and 1.25, which represents more aggressive weight boosting, achieves noticeable improvements. In other scenarios associated with less compact knowledge structures, e.g., complication, the difference is insignificant between the set of expansion concepts by our method and those by statistical expansion. As a result, the cumulative weights of the two set of expansion concepts are close to each other. For such scenarios, our experimental data suggests a more conservative weight boosting with $\alpha \in [0, 0.5]$.

Comparison with previous knowledge-based query expansion studies. Our research differs from most knowledge-based query expansion studies [Hersh et al.(2000), Plovnick and Zeng(2004), Guo et al.(2004)] in the baseline method used for comparison. Most existing studies only compare against a baseline generated by no query expansion. Such studies expand the synonyms, hypernyms and hyponyms of the original query concepts, and usually report an insignificant improvement [Guo et al.(2004)] or even degrading performance [Hersh et al.(2000)] compared to the no expansion method. In contrast, our study compares against statistical expansion which, in our experimental setup, has an observed improvement over no expansion by at least 5%.

In Aronson and Rindflesch's study [Aronson and Rindflesch(1997)], the researchers applied the UMLS Metathesaurus to automatically expand synonyms to the original query. In one particular setup, their approach achieved a 5% improvement over a previous study [Srinivasan(1996)] which applied statistical expansion on the same testbed. This result indicates the value of human knowledge in query expansion, and generally aligns with the observation in our experiments. We note that the difference between their research and ours is that their approach is limited to expanding synonyms only, and is not scenario-specific as we have presented in Section 1.

5 Knowledge Acquisition

The quality of our knowledge-based method largely depends upon the quality and completeness of the domain-specific knowledge source. The knowledge structure in the UMLS knowledge base is not specifically designed for scenario-specific retrieval. As a result, we discovered some frequently asked scenarios (e.g., etiology or complications of a disease) that are either undefined in UMLS, or defined but with incomplete knowledge. Therefore, we developed a methodology to acquire knowledge and to supplement the UMLS knowledge source. The methodology consists of the following two steps:

1. Acquire knowledge for undefined scenarios to supplement the UMLS knowledge source.
2. Refine the knowledge of the scenarios defined in the UMLS knowledge source (including the knowledge supplemented by Step 1).

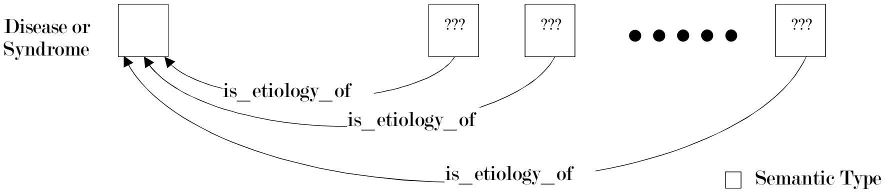


Figure 4: A sample template to acquire knowledge for previously undefined scenarios

5.1 Knowledge Acquisition Methodology

Knowledge Acquisition for Undefined Scenarios. For an undefined scenario, we present to medical experts an incomplete relationship graph as shown in Figure 4. Edges in this relationship graph are labeled with one of the undefined scenarios, e.g., "etiology." The experts will fill in the question marks with existing UMLS semantic types that fit the relationship. For example, because viruses are related to the etiology of a wide variety of diseases, the semantic type "Virus" will replace one of the question marks in Figure 4. This new relationship graph (etiology of diseases) will be appended to the UMLS Semantic Network, and can be used for queries with the "etiology" scenario.

Knowledge Refinement Through Relevance Judgments. A relationship graph for a given scenario (either previously defined by UMLS or newly acquired from Step 1) may be incomplete in including all relevant Semantic Types. A hypothetical example of this incompleteness would be the missing relationship `treats` between `Therapeutic or Preventive Procedure` and `Disease or Syndrome`. Our basic idea in amending this incompleteness is to explore the "implicit" knowledge embedded in the relevance judgments of a standard IR testbed. Such a testbed typically provides a set of benchmark queries and for each query, a pre-specified set of relevant documents. To amend the knowledge structure for a certain scenario, e.g., `treatment`, we focus on sample queries that are specific to this scenario, e.g., `keratoconus treatment`. We then study the content of documents that are marked as relevant to these queries. From the content, we can identify concepts that are directly relevant to the query's scenario, e.g., `treatment`. If the semantic type for those concepts are missing in the knowledge structure, we can then refine the knowledge structure by adding the corresponding semantic types. For example, let us consider a hypothetical case where the type `Therapeutic or Preventive Procedure` is missing in the knowledge structure of Figure 2. If by studying the sample query `keratoconus treatment`, we identify quite a few "`Therapeutic or Preventive Procedure`" concepts appearing in relevant documents such as `penetrating keratoplasty` and `epikeratoplasty`, we are then able to identify `Therapeutic or Preventive Procedure` as a relevant semantic type and append it to Figure 2.

Given that a typical benchmark query has a long list of relevant documents, it is labor-intensive to study the content of every relevant document. One way to accelerate this process is to first apply an incomplete knowledge structure to perform knowledge-based query expansion and perform retrieval tests based on such expansion. An incomplete knowledge structure leads to an “imperfect” query expansion, which in turn, fails to retrieve certain relevant documents to the top of the ranked list. Comparing this ranked list with the gold standard and identifying the missing relevant documents will give us pointers to determine the incomplete knowledge. For example, failure to include `Therapeutic or Preventive Procedure` in the knowledge structure in Figure 2 prevents us from expanding concepts such as `penetrating keratoplasty` to the sample query of `keratoconus, treatment`. As a result, documents with a focus on `penetrating keratoplasty` will be ranked unfavorably low. After we identify such documents, we can discover the missing expansion concepts contributing to the low rankings and refine the knowledge structure as we have just described.

5.2 Knowledge Acquisition Process

We chose the 57 scenario-specific queries (Table 4) in the OHSUMED testbed to apply our proposed knowledge-acquisition method because of the following considerations:

- The OHSUMED queries are collected from physicians treating patients in a clinical setting. Therefore, the OHSUMED query scenarios should be representative in healthcare, and the knowledge acquired from these scenarios should be broadly applicable.
- The knowledge-acquisition methodology also requires exploring relevance judgments for a set of benchmark queries. OHSUMED is the largest testbed for medical free-text retrieval that has relevance judgments for knowledge refinement.

We have identified 12 OHSUMED scenarios whose knowledge structures are missing in UMLS. We applied the two-step knowledge-acquisition method to acquire the knowledge structures for these 12 undefined scenarios and to refine the knowledge structures for all scenarios. During the first step of the acquisition process, we interviewed two intern doctors with M.D. degrees at the UCLA School of Medicine. During the interview, we first described the meaning of the relationship graphs as seen in Figure 4. Afterwards, we 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 from one expert to another until they reached a consensus for each scenario. For the second step of knowledge acquisition, we performed retrieval tests on the OHSUMED testbed using both queries expanded by the knowledge-based method and the method of expanding all statistically-related concepts. We focused on 12 queries where the statistical method outperforms the knowledge-based method in terms of the precision in top-10 results. We further applied the method presented in the

Scenarios	# of semantic types defined in UMLS	# of semantic types acquired from experts	# of additional semantic types through knowledge refinement	Total # of semantic types after knowledge acquisition
treatment of a disease	3	N/A	1	4
diagnosis of a disease	5	N/A	2	7
prevention of a disease	3	N/A	0	3
differential diagnosis of a symptom/disease	N/A	10	4	14
etiology of a disease	N/A	40	1	41
risk factors of a disease	N/A	40	2	42
complications of a disease/medication	N/A	15	0	15
pathophysiology of a disease	N/A	56	0	56
prognosis of a disease	N/A	15	2	17
epidemiology of a disease	N/A	13	0	13
research of a disease	N/A	28	0	28
organisms of a disease	N/A	7	0	7
criteria of medication	N/A	26	0	26
when to perform a medication	N/A	5	6	11
preventive health care for a type of patients	N/A	10	2	12

Table 18: Knowledge acquisition results

previous section to study the content of these top-ranked documents and augmented the knowledge structure for the corresponding scenario with appropriate semantic types.

5.3 Knowledge Acquisition Results

The acquisition results are shown in Table 18. Due to space constraints, we only provide a statistical summary of the results. Appendix A presents the results in full detail.

The scenarios in the first three rows, i.e., `treatment`, `diagnosis` and `prevention`, are originally defined by UMLS. The first column in these rows shows the number of semantic types marked as relevant for each scenario (i.e., the number of semantic types that experts have filled into the blank rectangles of Figure 4). The second column for these rows is "N/A" because there was no need to acquire knowledge structure from domain experts for these scenarios. The third column shows the number of semantic types added during knowledge refinement (the second step of knowledge acquisition). For example, for the `diagnosis` scenario two additional semantic types, `Laboratory or Test Result` and `Biologically Active Substance` were added because of the study on `Query #97: Iron deficiency anemia, which test is best`. These two semantic types were added because the absence of these two types has prevented the knowledge-based method from expanding two critical concepts into the original query: `serum ferritin` and `fe iron`, each belonging to one of the two semantic types. From the relevance judgment set, we noted that missing these two concepts leads to the low ranking of three relevant documents that heavily use these two concepts.

Starting from the fourth row, we list the scenarios for which we need to acquire knowledge structure from domain experts. The first column for these scenarios is

"N/A" because these scenarios are originally undefined in UMLS. The second column shows the number of semantic types that experts have filled into the structure template of Figure 4. The third column shows the number of additional semantic types from knowledge refinement (the second step of knowledge acquisition), and the last column shows the total number of semantic types after knowledge acquisition.

The proposed knowledge-acquisition method on the OHSUMED testbed has shown to be efficient and effective. We finished communicating with domain experts to acquire the knowledge structures for the 12 scenarios in less than 20 hours, and spent an additional 20 hours to refine the knowledge structure by exploring the relevance judgments. We applied the augmented knowledge source in our knowledge-based query expansion experiments. The augmented knowledge was shown to be effective in helping improve the retrieval performance of the knowledge-based method over the statistical expansion method.

6 Study of The Relevancy of Expansion Concepts by Domain Experts

Through experiments on the two standard medical text retrieval testbeds, we have observed that under most retrieval settings knowledge-based query expansion outperforms statistical expansion. Our conjecture is that knowledge-based query expansion selects more specific expansion concepts to the original query's scenario than statistical expansion does. To verify this conjecture, we have asked domain experts to manually evaluate the relevancy of expansion concepts.

The basic idea for this study is the following: For each query in a given retrieval testbed, we apply the two query expansion methods to generate two sets of expansion concepts. We then prepare an evaluation form which inquires about the relevancy of each expansion concept to the original query. In this form, we present the query's text, and ask domain experts to judge the relevancy based on the query's scenario(s). For each concept we provide four scales of relevancy: *relevant*, *somewhat relevant*, *irrelevant*, or *do not know*. A concept will be marked as somewhat relevant if the concept is indirectly related to the original query or conditionally relevant in certain clinical cases. We blind the method used to generate each concept. In doing so, we reduce bias that an expert might have towards a particular method.

To implement this idea, we chose the 57 scenario-specific queries in the OHSUMED testbed. We applied the two expansion methods and derived 40 expansion concepts from each method with the highest weights. We presented the evaluation form consisting of these concepts to three medical experts who are intern doctors at the UCLA School of Medicine. We asked them to make judgments only on those queries that belong to their area of expertise, e.g., oncology, urology, etc. On average, each expert judged the expansion concepts for 15 queries. Thus, for each expansion method, we obtained 1,600 expansion concepts classified as one of the four categories.

Figure 5 and Figure 6 present a summary of the results from this human subject

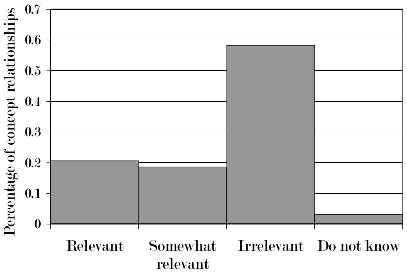


Figure 5: Relevancy of expansion concepts created by statistical expansion

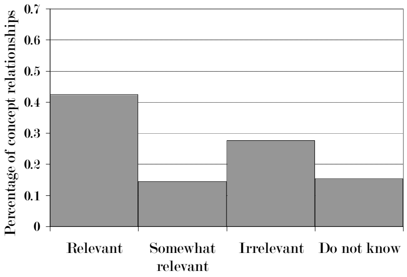


Figure 6: Relevancy of expansion concepts created by knowledge-based expansion

study. For the expansion concepts derived from each method, we summarized the results into a histogram. The bins of this histogram are the four scales of relevancy. We note that 56.9% of the expansion concepts derived by the knowledge-based method are judged as either *relevant* or *somewhat relevant*, whereas only 38.8% of expansion concepts by statistical expansion are judged similarly. This represents a 46.6% improvement. This result validates that knowledge-based query expansion derives more relevant expansion concepts to the original query's scenario(s) than those by statistical expansion, and thus yields improved retrieval results for scenario-specific queries.

7 Conclusion

Scenario-specific queries represent a special type of query that is frequently used 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. We have made the following contributions:

- We have developed a methodology that 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 query. Appending such identified concepts to the query results in scenario-specific expansion.
- We have developed an efficient and effective methodology for knowledge acquisition to supplement and refine the knowledge source.
- We have performed extensive experimental evaluation of the retrieval performance of knowledge-based query expansion by comparing with that of statistical expansion. Our experimental studies reveal that:
    - Knowledge provided by UMLS is useful in creating scenario-specific query expansion, leading to over 5% of improvements over statistical expansion in the majority of cases studied. Such improvements are significant since

statistical expansion outperforms the no-expansion method by at least 5% in our experimental setup.

- Since knowledge-based expansion tends to expand less terms into the original query, boosting the weights of these terms is necessary to generate improvements over the statistical method.
- Because the knowledge structures defined for different query scenarios exhibit different characteristics, the performance improvements of the knowledge-based expansion method differ for these scenarios.

The focus of this research is to support scenario-specific queries in the medical domain. Scenario-specific queries can appear in other domains as well. In extending our research to other domains, we note that the quality of domain knowledge is important to the performance of our method. In certain domains where such knowledge is not readily available, the success of our approach depends on the knowledge acquisition process which is resource-intensive. This represents a limitation of our current approach in terms of extensibility across different domains, and is a worthy topic for future research.

Acknowledgement

This research is supported in part by NIC/NIH Grant #4442511- 33780. We are grateful to Dr. Andrew Chen and Wei Liu from the UCLA School of Medicine for providing their domain knowledge during the knowledge acquisition task. We thank Dr. Blaine Kristo and his colleagues for their efforts in evaluating the relevancy of expansion terms. We also thank our colleagues in the UCLA Department of Radiology Sciences, Dr. Hooshang Kangarloo, Dr. Denise Aberle, Dr. Suzie El-Saden, for stimulating discussions and insightful comments. Finally, we thank Nancy Wilczynski and Dr. Brian Haynes from the Health Information Research Unit (HIRU) at McMaster University for sharing their valuable dataset for our experimental verification of the effectiveness of the knowledge-based expansion methodology.

References

[Aho and Corasick(1975)] A.V. Aho and M.J. Corasick. Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, 18(6):330–340, 1975.

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

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

[Buckley et al.(1994)] 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.

[Buckley et al.(1995)] 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.

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

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

[Efthimiadis and Biron(1993)] 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.

[Ely et al.(1999)] 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.

[Ely et al.(2000)] 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.

[Guo et al.(2004)] Y. Guo, H. Harkema, and R. Gaizauskas. Sheffield university and the trec 2004 genomics track: Query expansion using synonymous terms. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC-13)*, 2004.

[Haynes et al.(1990)] 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.

[Hersh et al.(1994)] 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.

[Hersh et al.(2000)] 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.

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

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

[Liu and Chu(2006)] Z. Liu and W.W. Chu. Knowledge-Based Query Expansion to Support Scenario-Specific Retrieval of Medical Free Text. Technical Report #060019, Computer Science Department, UCLA, `ftp://ftp.cs.ucla.edu/tech-report/2006-reports/060019.pdf`, 2006.

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

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

[Montori et al.(2003)] V.M. Montori, N.L. Wilczynski, D. Morgan, and R.B. Haynes. Systematic reviews: A cross-sectional study of location and citation counts. *BMC Medicine*, 1(2), 2003.

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

[Plovnick and Zeng(2004)] R.M. Plovnick and Q.T. Zeng. Reformulation of consumer health queries with professional terminology: a pilot study. *Journal of Medical Internet Research*, 6(3), 2004.

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

[Robertson et al.(1994)] 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.

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

[Salton and Buckley(1988)] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

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

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

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

[Tse and Soergel(2003)] T. Tse and D. Soergel. Exploring medical expressions used by consumers and the media: An emerging view of consumer health vocabularies. In *Proceedings of AMIA Annual Symp 2003*, 2003.

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

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

[Wilczynski and Haynes(2003)] N.L. Wilczynski and R.B. Haynes. Developing optimal search strategies for detecting sound clinically sound causation studies in MEDLINE. In *Proceedings of AMIA Annual Symp 2003*, 2003.

[Wilczynski et al.(2001)] 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.

[Wong et al.(2003)] S.-L. Wong, N.L. Wilczyski, R.B. Haynes, and R. Ramkissoonsingh. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. In *Proceedings of AMIA Annual Symp 2003*, 2003.

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

[Zeng et al.(2002)] Q. Zeng, S. Kogan, N. Ash, R.A. Greenes, and A.A. Boxwala. Characteristics of consumer terminology for health information retrieval. *Methods in Information in Medicine*, 41(4):289–298, 2002.

[Zou et al.(2003)] 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.

Appendix A. Knowledge Acquisition Results

We first consulted medical experts to acquire the set of semantic types relevant to a scenario that is previously undefined by UMLS (e.g “etiology of disease”). Table 19 lists the acquisition results. Due to space limit, we only provide the ID of each semantic type. The definition can be found in the “SRDEF” table of UMLS. We further performed knowledge refinement through relevance judgements and present the refinement results in Table 20. The second column in the figure shows the additional semantic types added to the corresponding scenario from this step.

Scenario	Set of relevant semantic types by consulting experts (ID's only)
differential diagnosis of a disease	T059, T060, T097, T121, T184, T046, T047, T048, T049, T191
etiology of a disease	T004, T005, T006, T007, T009, T031, T073, T074, T075, T103, T104, T105, T106, T107, T108, T109, T110, T111, T112, T113, T114, T115, T116, T118, T119, T120, T121, T122, T123, T124, T125, T126, T127, T128, T129, T130, T131, T167, T168, T192, T053, T054, T055, T047, T048, T191, T049, T190, T019, T020, T037
risk factors of a disease	T004, T005, T006, T007, T009, T031, T073, T074, T075, T103, T104, T105, T106, T107, T108, T109, T110, T111, T112, T113, T114, T115, T116, T118, T119, T120, T121, T122, T123, T124, T125, T126, T127, T128, T129, T130, T131, T167, T168, T192, T053, T054, T055, T047, T048, T191, T049, T190, T019, T020, T037
complications of a disease/medication	T033, T034, T184, T059, T060, T061, T047, T048, T190, T019, T020, T054, T055, T080, T081
pathophysiology of a disease	T062, T059, T039, T040, T041, T042, T043, T044, T045, T046, T047, T048, T191, T049, T050, T018, T021, T023, T024, T025, T026, T028, T190, T019, T020, T109, T110, T111, T112, T113, T114, T115, T116, T118, T119, T120, T121, T122, T123, T124, T125, T126, T127, T128, T129, T192, T033, T034, T184, T085, T086, T087, T088, T169, T022, T059
prognosis of a disease	T033, T034, T184, T059, T060, T061, T047, T048, T190, T019, T020, T054, T055, T080, T081
epidemiology of a disease	T083, T097, T098, T099, T100, T101, T102, T002, T003, T004, T005, T006, T007
research of a disease	T062, T063, T109, T110, T111, T112, T113, T114, T115, T116, T118, T119, T123, T124, T125, T126, T127, T128, T129, T192, T085, T086, T087, T088, T004, T005, T006, T007
organisms for a disease	T001, T002, T003, T004, T005, T006, T007
criteria of medication	T033, T034, T184, T059, T060, T109, T110, T111, T112, T113, T114, T115, T116, T118, T119, T123, T124, T125, T126, T127, T128, T129, T192, T190, T019, T020
when to perform a medication	T059, T060, T033, T034, T184
preventive health care for a type of patient	T059, T053, T054, T055, T056, T064, T065, T124, T127, T129, T080, T169

Table 19: The set of semantic types relevant to each scenario, results acquired by consulting medical experts

Scenario	Set of relevant semantic types appended during knowledge refinement
treatment of a disease	T093
diagnosis of a disease	T034, T123
prevention of a disease	n/a
differential diagnosis of a disease	T034, T123, T031, T082
etiology of a disease	T059
risk factors of a disease	T059, T034
complications of a disease/medication	n/a
pathophysiology of a disease	n/a
prognosis of a disease	T169, T025
epidemiology of a disease	n/a
research of a disease	n/a
organisms for a disease	n/a
criteria of medication	n/a
when to perform a medication	T046, T047, T048, T049, T191, T023
preventive health care for a type of patient	T080, T169

Table 20: The set of semantic types relevant to each scenario, acquired from knowledge refinement by exploring relevance judgements