Information Retrieval, 5, 323–351, 2002

Exploiting Manual Indexing to Improve Collection Selection and Retrieval Effectiveness

JAMES C. FRENCH* french@cs.virginia.edu
Department of Computer Science, University of Virginia, Charlottesville, VA, USA

ALLISON L. POWELL† apowell@cnri.reston.va.us
Corporation for National Research Initiatives, Reston, VA, USA

FREDRIC GEY gey@ucdata.berkeley.edu
NATALIA PERELMAN‡
UC Data Archive & Technical Assistance, University of California, Berkeley, CA, USA

Received February 26, 2002; Revised May 22, 2002; Accepted May 23, 2002

Abstract. Vocabulary incompatibilities arise when the terms used to index a document collection are largely unknown, or at least not well-known to the users who eventually search the collection. No matter how comprehensive or well-structured the indexing vocabulary, it is of little use if it is not used effectively in query formulation. This paper demonstrates that techniques for mapping user queries into the controlled indexing vocabulary have the potential to radically improve document retrieval performance. We also show how the use of controlled indexing vocabulary can be employed to achieve performance gains for collection selection. Finally, we demonstrate the potential benefit of combining these two techniques in an interactive retrieval environment. Given a user query, our evaluation approach simulates the human user's choice of terms for query augmentation given a list of controlled vocabulary terms suggested by a system. This strategy lets us evaluate interactive strategies without the need for human subjects.

Keywords: entry vocabulary, collection selection, interactive retrieval, query augmentation

1. Introduction

In recent years there has been renewed interest in document collections that have been manually indexed with terms assigned by human indexers. Index terms can come from controlled or uncontrolled vocabularies and can be assigned by either authors or professional indexers. In this work, we are investigating query expansion where one or more terms are drawn from a controlled vocabulary, the indexing vocabulary used for manual indexing. In our terminology Q, the *original query*, is expanded by the addition of these term(s) to become Q', the *augmented query*.

*This work supported in part by DARPA contract N66001-97-C-8542 and NASA Grants NAG5-8585 and NAG5-9747.

†This work supported in part by DARPA contract N66001-97-C-8542 and SPAWAR cooperative agreement N66001-00-2-8902. The work reported here was performed when the author was at the University of Virginia.

‡This work supported in part by DARPA Contract N66001-97-8541.

We are investigating the effects of query augmentation in two arenas. We consider query augmentation for a straightforward document retrieval scenario. We also consider query augmentation in a distributed or multi-collection environment. For the latter case, we study the effects of query augmentation for both collection selection and for multi-collection document retrieval. Our goal is to investigate two main questions:

1. How does the use of augmented queries for collection selection compare to the use of the original free text queries?
2. What is the effect when augmented queries are used for document retrieval?

In the discussion that follows, we will cover a number of points. We will discuss related work in query augmentation and in collection selection. We will describe two multi-collection test environments based on the OHSUMED (Hersh et al. 1994) test collection and discuss features of those test environments. We will present two concrete approaches to query augmentation that allow us to tap into the controlled vocabulary terms (Medical Subject Headings or MeSH terms) that have been manually assigned to the documents in the OHSUMED test collection. Given these approaches to query augmentation, we will present results that measure their effects on both collection selection and document retrieval.

We restate the general questions from above as a set of hypotheses to focus our discussion.

Hypothesis 1. Augmented queries will be more effective for collection selection than the original queries. Adding more MeSH headings will improve collection selection results.

Hypothesis 2. The benefits of using augmented queries for collection selection will translate to superior document retrieval results, even when the original queries are used for document retrieval.

Hypothesis 3. Augmented queries will outperform the original queries for document retrieval.

2. System configurations

Here we briefly overview the possible configurations for an information retrieval system employing query augmentation and/or collection selection. We hope this will help the reader visualize the space of options that we are considering.

Figure 1 is a schematic view of the possible scenarios involving information retrieval systems using one or both of query augmentation and collection selection. We show a generic query augmentation component (QA) which takes a query Q as input and produces, by some strategy, an augmented query Q' as output. We also show a generic collection selection component (CS) which takes a query as input and produces a list of collections to search as output.

We can begin to understand the effects of these components by comparing various system configurations. That is the basis of our experimental methodology. Here we review the

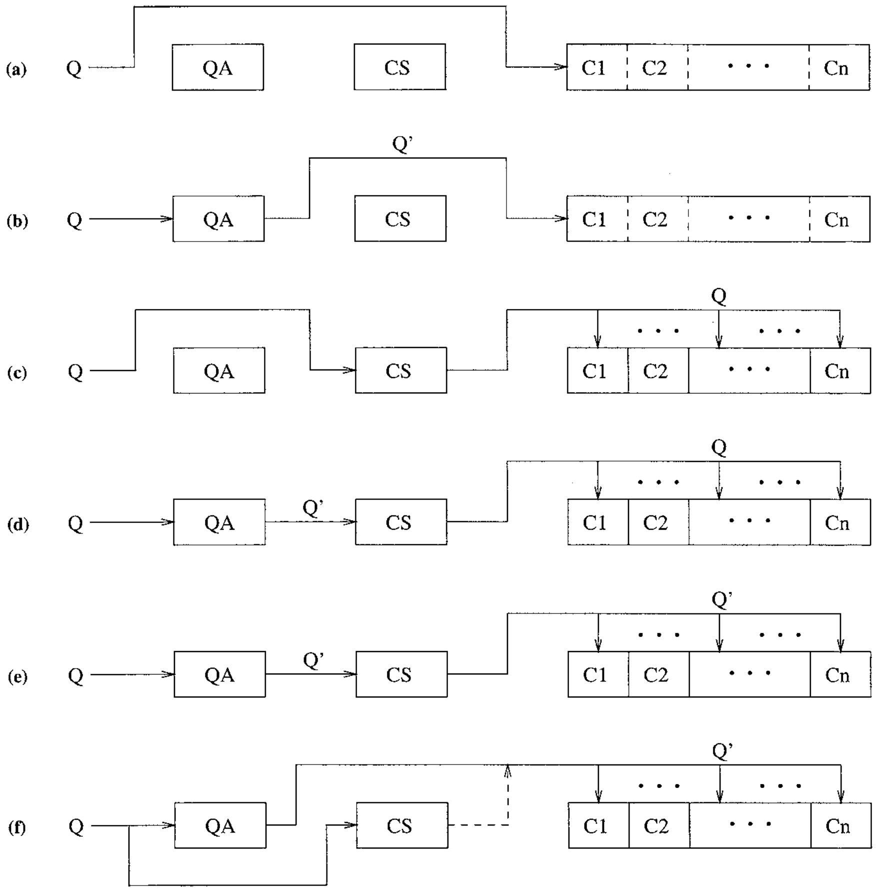


Figure 1. A schematic diagram of all the configurations for query augmentation (QA) and collection selection (CS) used in this study. Q denotes the original query and Q' denotes the augmented query. C_i denotes the ith collection. In (a) and (b) dashed lines are used to denote the union of the n collections. In (c)–(f) solid lines indicate distinct collection boundaries.

configurations used in our study. In later sections we will relate our results back to this schematic to provide a context for interpreting those results.

Figure 1(a) is the default configuration. A query is submitted to the *union collection*, $\mathcal{C} = \cup C_i$; the search engine technology is left unspecified. Figure 1(b) introduces a QA component and uses an augmented query to search the union collection. In figure 1(c) a collection selection component is used alone and the original query is directed to the selected collections shown here as subsets of the union collection. Figure 1(d) and (e) use both components and use the augmented query for selection. They differ only in what query is sent to the selected subsets; the original query is used in (d) while the augmented query is used in (e). Figure 1(f) shows the case where the original query is used for selection but the augmented query is executed at the selected sites.

The diagram exposes many of the aspects of the experimental setting that have to be selected, controlled or varied. Specifically, we have to make choices for:

1. the query augmentation strategy QA;
2. the collection selection strategy CS; and
3. the search engine technology.

We discuss our choices for these parameters further in Section 6.7.

3. Background and related work

Our work is focused on augmenting queries with terms drawn from a controlled vocabulary to enhance collection selection and document retrieval. Manual indexing is a labor intensive activity that provides enormous potential for improving retrieval performance. Our work seeks to take advantage of such manually acquired terms for both collection selection and document retrieval. A preliminary version of this work was reported in French et al. (2001). Here we elaborate and expand that work and corroborate the findings with another set of experiments.

3.1. Manual indexing

The NTCIR collection of Japanese-English scientific abstracts from 65 scientific societies of Japan presents an example of author-assigned terms without vocabulary restriction. It has been utilized in the NTCIR evaluation of Japanese and Japanese-English text retrieval (Kando et al. 1999). The index terms are a useful source for cross-language information retrieval because authors have assigned keywords in both Japanese and English. The GIRT (German Information Retrieval Test) collection (Kluck and Gey 2001) consists of German language abstracts in the social sciences which have been indexed professionally using the vocabulary contained in the GIRT thesaurus (Schott 2000). The thesaurus is multi-lingual in German, English, and Russian. The OHSUMED collection (Hersh et al. 1994), which is the focus of experiments in this paper, consists of a strict subset of the MEDLINE medical domain abstracts, with index terms assigned by professional indexers from the MeSH thesaurus. The MeSH vocabulary has been translated into Spanish and has been utilized by Eichmann, Ruiz and Srinivasan for cross-language information retrieval (Eichmann et al. 1998). The INSPEC collection of scientific and engineering abstracts indexed with the INSPEC thesaurus provides a commercial example of this genre of document collections.

An interesting research question is whether the intellectual value-added of human indexing can provide leverage for improved information retrieval through mechanisms of query expansion, either automatically or as part of an interactive relevance feedback loop with a user involved in term selection. A simple term-matching approach to suggesting MeSH terms for medical searching was implemented in CITE (Doszkocs 1983), however no effectiveness results were reported. Shatz, Chen and colleagues have provided a design for interactive term suggestion from the INSPEC subject thesaurus and contrasted it to the alternative of co-occurrence lists (Schatz et al. 1996). Gey et al. (2001) have been studying the interactive suggestion of subject terms to users by probabilistic mapping between the user's natural language and the technical classification vocabularies through a methodology called Entry Vocabulary Indexes (EVIs) (Buckland et al. 1999, Gey et al. 2001).

When a controlled vocabulary thesaurus is utilized for indexing, a natural approach to query expansion is to add narrower terms to terms found in documents. Hersh and his colleagues have studied the effect of automatic narrower-term expansion for OHSUMED and concluded that while performance improves for some queries, overall performance declines (Hersh et al. 2000). This approach contrasts with the widely used technique of pseudo-relevance or "blind" feedback wherein the top documents of an initial ranking are mined for additional natural language terms to be added to the initial query. Both techniques have counterparts in interactive relevance feedback wherein either documents or suggested terms can be presented to the user who chooses which words, phrases, or terms are to be added to the query.

3.2. Collection selection

The problem of document retrieval in a multi-collection environment can be broken down into three major steps. First, given a set of collections that may be searched, the collection selection step chooses the collections to which queries will be sent. Next, the query is processed at the selected collections, producing a set of individual result-lists. Finally, those result-lists are merged into a single list of documents to be presented to a user.

A number of different approaches for collection selection using free-text queries have been proposed and individually evaluated (Callan et al. 1995, Fuhr 1999, Gravano et al. 1999, Hawking and Thistlewaite 1999, Meng et al. 1998, Yu et al. 1999, Yuwono and Lee 1997). Three of these approaches, *CORI* (Callan et al. 1995), *CVV* (Yuwono and Lee 1997) and *gGlOSS* (Gravano et al. 1999) were evaluated in a common environment by French et al. (Callan et al. 2000, French et al. 1998, 1999), who found that there was significant room for improvement in all approaches, especially when very few databases were selected. One of the goals of these experiments is to determine if the use of augmented queries can provide that improvement.

Other work has shown that improvements in collection selection performance can translate into improved document retrieval performance (Powell et al. 2000, Xu and Callan 1998). Of particular interest to us here is the work of Xu and Callan (1998) who noted that query expansion can improve collection selection performance. Xu and Callan studied query expansion using the general vocabulary of documents in the collections, however in this work we consider the effect of augmented queries using controlled vocabulary.

4. OHSUMED-based test environment

All of the experiments reported here were conducted using specific organizations of the documents found in the OHSUMED test collection. The OHSUMED collection, constructed and described by Hersh et al. (1994), contains bibliographic entries and abstracts for 348,566 MEDLINE medical articles. A set of 106 queries and corresponding relevance judgements are provided. Of the 348,566 entries, 233,445 have abstracts and 348,543 have had MeSH controlled vocabulary entries manually assigned.

The manually-assigned MeSH headings make the OHSUMED collection useful for our study of augmented queries; however, we are interested in the effect of augmented queries on both document retrieval and collection selection. Because we are interested in distributed information retrieval in general and collection selection in particular, we needed to organize the OHSUMED documents into multiple collections. We chose to organize the documents into testbeds using two different strategies termed *journal-based* and *subject-based*. Each is discussed below.

Journal-based decomposition: The documents were organized according to journal of publication to create a multi-collection test environment. This yielded 263 collections and provides us with a test environment that has a topical organization (i.e. many of the journals focus on specific medical subfields).

Subject-based decomposition: The documents were organized by subject. We tried to group the journals into subject categories that would be of interest to or read by specific medical subdisciplines. We were aided in this task by medical informatics researchers. The decomposition resulted in 48 collections. The explicit decomposition, that is, the assignment of journals to subject categories is shown in Appendix B.

A number of choices had to be made concerning the actual parsing and handling of the OHSUMED data. These were practical decisions to assure uniform handling of the data for our experiments. These issues are briefly covered in Appendix A.

There are a number of interesting features of the OHSUMED collection and of our organization of the OHSUMED documents into multi-collection environments. First, we will discuss features of the queries and relevance judgements, then we will discuss the distribution of relevant documents among our journal-based and subject-based collections.

4.1. Queries and relevance judgements

The OHSUMED test collection is accompanied by 106 queries and two sets of relevance judgements.

The queries are fielded and contain two types of information. One field contains a direct statement of information need, while a second provides biographical information about the patient whose condition prompted the query. In our experiments we used only the statement of information need as the original query. An example query, Query 83 from the OHSUMED testbed, is shown in figure 2. Again, only the statement of information need is shown.

There are two sets of relevance judgements associated with the queries. The documents were judged on a ternary scale—“definitely relevant”, “possibly relevant” and “not relevant”. For our experiments, we used a binary scale for relevance judgements and counted “possibly relevant” documents as “not relevant”. For 5 queries, there are no documents that were judged “definitely relevant”. We excluded those queries and use the remaining 101 queries for our experiments.

Query:
Infections in renal transplant patients
Suggested MeSH terms:
Kidney Transplantation
Kidney/TR

Figure 2. Query 83 from the OHSUMED testbed shown with two suggested MeSH terms.

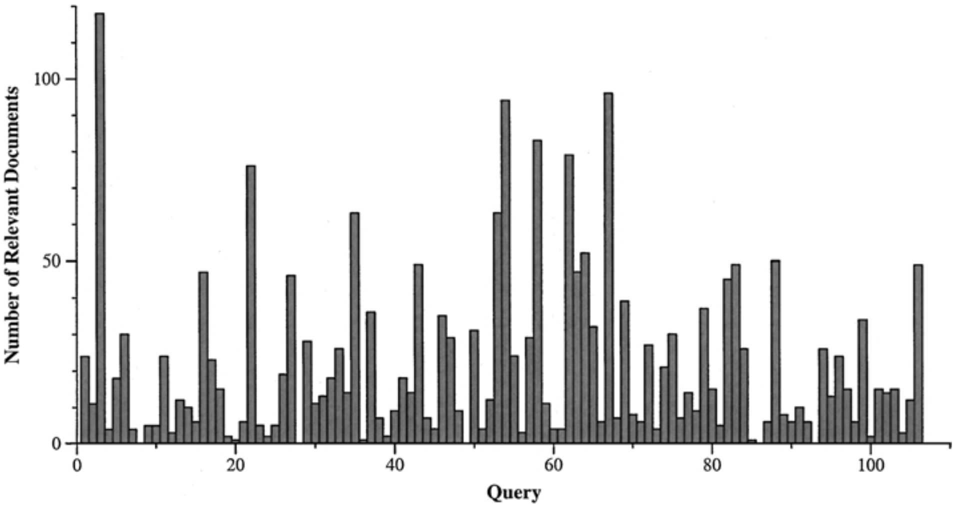


Figure 3. Number of relevant documents per query.

4.2. Distribution of documents and relevant documents

The OHSUMED test collection, and our journal-based and subject-based organizations of the documents into multi-collection testbeds make an interesting and sometimes challenging test environment. Despite the specialized vocabulary of the documents and queries, the environment can be challenging due to the relatively small number of relevant documents overall. On average, there are only 22.3 relevant documents per query, with a minimum of 1 and a maximum of 118 (figure 3).

Our choice of organizing the documents by publishing journal (subject category) resulted in a skewed distribution of documents among collections. On average, there are 1,325 (7,262) documents per collection with a minimum of 3 (399) and a maximum of 12,654 (55,384). Most challenging from a collection-selection point of view is the fact that despite the skew in the distribution of documents, the *relevant* documents tend to be very evenly distributed across the collections for many queries (figure 4). For the 263 collection journal-based decomposition, of the 101 queries under consideration, 45 have two or fewer relevant documents in the collection containing the *most* relevant documents. Only 21 queries have an average of two or more relevant documents per collection. This type of scenario has been shown to be particularly challenging for collection selection (French and Powell 2000). The

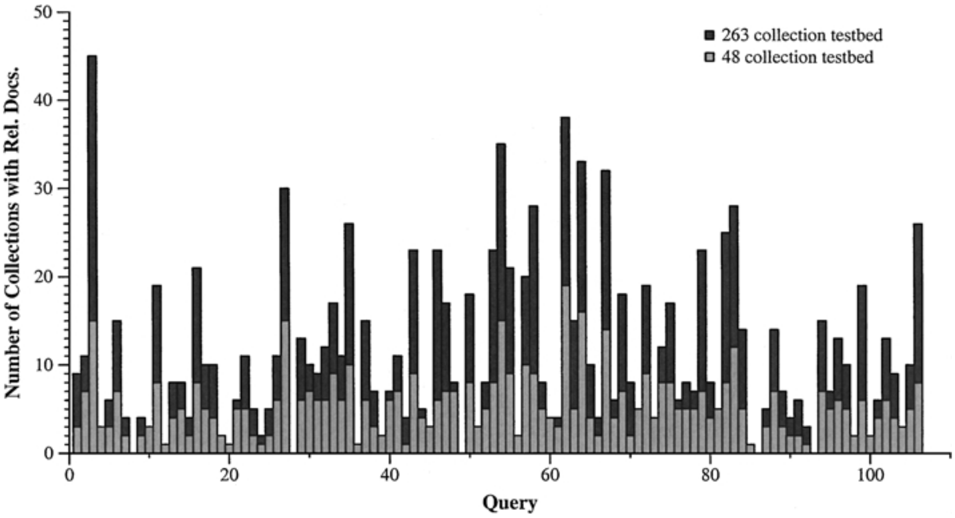


Figure 4. Number of collections containing at least one relevant document.

distribution of relevant documents is slightly more skewed in the 48 collection subject-based decomposition. However, the test environment is still challenging. 22 queries have two or fewer relevant documents in the collection containing the most relevant documents and only 73 queries have an average of two or more relevant documents per collection.

Although both decompositions are challenging testbeds, we would expect the 48 collection testbed to be somewhat less challenging than the 263 collection testbed for a number of reasons. The distribution of relevant documents in collection is more skewed, potentially making the collection selection task easier. Plus, due to the smaller number of collections, we would expect a greater proportion of collections to have relevant documents in the 48 collection decomposition. For example, consider Query 3 in figure 3. Relevant documents occur in 31% (15/48) of the collections for the 48 collection testbed as compared with just 17% (45/263) of the collections in the 263 collection testbed.

5. Query augmentation approaches

Query augmentation is achieved in one of two ways: (1) automatic query expansion; or (2) term suggestion. We are investigating the latter approach in which we use an entry vocabulary index (EVI) to suggest MeSH terms that are appropriate for the original query. In our experiments, we use two query augmentation approaches. One approach augments the queries with terms suggested by an existing term suggestion mechanism, referred to here as an Entry Vocabulary Index (EVI). The second approach augments the queries with the MeSH terms most frequently assigned to relevant documents, a strategy referred to later as *RBR-EVI*. Figure 2 shows an example of EVI suggested terms for Query 83 of the OHSUMED testbed. The qualifier `TR` in the MeSH heading `Kidney/TR` is the topical subheading "transplant."

Table 1. Contingency table from words/phrases to classification.

	C	$\neg C$
t	a	b
$\neg t$	c	d

5.1. Entry vocabulary indexes

Construction of Entry Vocabulary Indexes rests upon three basic components: (1) a sufficiently large training set of documents that have been manually indexed with a metadata classification or thesaurus; (2) software and algorithms to develop probabilistic mappings between words in the document text and metadata classifications; and (3) software to accept search words/phrases and return classifications. For this research we utilized the entire collection of OHSUMED documents and assigned MeSH terms for our training set. Research on relevance feedback has suggested that collection-specific term suggestion can be even more effective (Gauch et al. 1999). We plan to investigate collection-specific EVIs in future work.

The final stage to creation of an Entry Vocabulary Index is the use of a maximum likelihood weighting associated with each text term and each subject heading. One constructs a two-way contingency table for each pair of terms t and classifications C as shown in Table 1.

Where a is the number of document titles/abstracts containing the word or phrase and classified by the classification; b is the number of document titles/abstracts containing the word or phrase but not classified by the classification; c is the number of titles/abstracts not containing the word or phrase but is classified by the classification; and d is the number of document titles/abstracts neither containing the word or phrase nor being classified by the classification.

The association score between a word/phrase t and a classification C is computed following Dunning (1993)

$$\begin{aligned} W(C,t) = 2[&\log L(p_1, a, a+b) + \log L(p_2, c, c+d) \\ &- \log L(p, a, a+b) - \log L(p, c, c+d)] \end{aligned}$$

where

$$\log L(x, n, k) = k \cdot \log(x) + (n-k) \cdot \log(1-x)$$

and $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, and $p = \frac{a+c}{a+b+c+d}$.

5.2. RBR-EVI

We are interested in gauging the *potential* of query augmentation in this environment. Therefore, we constructed an oracle, referred to here as *RBR-EVI*, to select MeSH terms for query augmentation. The premise behind *RBR-EVI* is that the best MeSH terms with which to augment a query are the principal[1] MeSH terms that have been assigned to the greatest number of documents relevant to that query.

For each query, we examine the set of relevant documents for that query and maintain a histogram of MeSH terms assigned to those documents. We sort the MeSH terms in decreasing order of the number of relevant documents to which they were assigned to create a list of MeSH terms from which to choose. For our experiments, we add the top-ranked 1, 2 and 3 MeSH terms to create *RBR-EVI* augmented queries. This approach is not necessarily optimal; for example, if the second-ranked term co-occurs frequently with the top-ranked term then adding the second-ranked term may not improve performance for that query. However, *RBR-EVI* does suggest very good MeSH terms.

We re-iterate that the *RBR-EVI* approach to augmenting queries is an attempt to gauge the potential of query augmentation. This approach can only be employed when relevance judgements are available.

5.3. Simulating user interaction

Term suggestion is an interactive technique in which the searcher is presented with a list of terms (in our case ranked) from which to choose appropriate MeSH terms to add to the original query. Operationally the original query is presented to the EVI and a ranked list of suggested terms is displayed to the searcher. To simulate the user interaction we are immediately faced with the decision of which terms to select. Because we present a ranked list of say n terms, it is tempting to simply augment the query with the first k suggested terms on the assumption that they are somehow the "best." But, this is not how humans approach the task. In particular, a human searcher would scan the entire list (provided it is of reasonable size) and pick the best terms to add to the query based on an internalized information need. Moreover, if told to augment a query with k terms, a human would interpret that to mean *at most k* terms, preferring to add fewer or none at all when the suggested terms did not look promising.

These observations lead to our strategy of simulating an expert user.[2] We have a concrete EVI instance that we are evaluating. For the testbed, we also have an oracle, *RBR-EVI* that largely represents an upper bound on achievable performance. Our strategy is to combine them to simulate a knowledgeable searcher. We do so as follows. First the query is presented to the EVI and a list of terms is suggested. That list is then intersected with the *RBR-EVI* term suggestions. The rationale is that if the *RBR-EVI* terms appear among the EVI suggestions, then those are precisely the terms the knowledgeable user would select for query augmentation. Because the *RBR-EVI* contains the $k = 3$ best MeSH terms, our simulated interaction (SI) adds at most 3 MeSH terms to the original query. Our approach is similar to the one used by Harman (1988) for query expansion using the general vocabulary of a collection.

To summarize, the SI approach combines an oracle and an algorithm (e.g., term suggestion, collection selection, etc.) to simulate "good" choices made by a knowledgeable user. The assumptions underlying the SI approach are reasonable. The technique allows us to simulate interactive retrieval techniques in a laboratory setting and provides an alternative means of gauging the effectiveness of interactive techniques without the need for costly user studies. We demonstrate the use of this technique in Section 7.

Table 2. Top 15 ranked MeSH terms suggested by EVI for Query 84 of the OHSUMED testbed.

Query: Theophylline uses-chronic and acute asthma

Rank	MeSH term	MeSH qualifier
1.	**Asthma/DT**	**Drug therapy**
2.	Asthma/PP	Physiopathology
3.	Theophylline/PK	Pharmacokinetics
4.	Theophylline/PD	Pharmacology
5.	Myocardial Infarction/DT	Drug therapy
6.	Asthma/EP	Epidemiology
7.	**Theophylline/TU**	**Theraputic use**
8.	Asthma/TH	Therapy
9.	Asthma/ET	Etiology
10.	Asthma/MO	Mortality
11.	Theophylline/AD	Administration
12.	Theophylline/BL	Blood
13.	Asthma/DI	Diagnosis
14.	Asthma/CO	Complications
15.	**Lung diseases, obstructive/DT**	**Drug therapy**

Bold terms overlap with *RBR-EVI*. The expansions of the two character MeSH topical subheadings are shown in the third column.

5.4. A concrete example

Now that we have discussed our query augmentation approach, it will be instructive to illustrate our approach with a concrete example. The information need expressed in Query 84 of the OHSUMED testbed is "*theophylline uses—chronic and acute asthma*." An examination of the relevant documents makes it clear that the physician was seeking guidance as to the therapeutic use of theophylline for asthma patients.

Table 2 shows the top 15 ranked MeSH headings suggested by our EVI approach. The three boldfaced entries are the top three ranked MeSH terms taken from *RBR-EVI*. (The qualifier expansions are also shown in the table for clarity.) By hypothesis our SI approach has identified the three MeSH terms that a knowledgeable searcher would select when presented with this list of suggested terms given the specific information need.

As will be discussed later, our query augmentation experiments consider adding up to three MeSH terms to the original query. Details of the outcome of these experiments are given in Section 7. Here we consider the effect of adding the three MeSH terms highlighted in Table 2 to Query 84. This is merely intended to show the potential of the technique. Table 3 compares the performance of the augmented query against that of the original query. The performance metric used is precision at *n* documents retrieved. Precision is simply the ratio of the number of relevant documents retrieved to the number of documents retrieved, *n* in this case. As Table 3 shows, the augmented query achieves precision greater than or equal to

Table 3. Precision at n documents for original OHSUMED query 84 and for query augmented by SI strategy using MeSH terms highlighted in Table 2.

OHSUMED query 84 precision results		
n	Original query	Simulated interaction
5	0.4000 (2)	0.4000 (2)
10	0.3333 (3)	0.4000 (4)
15	0.2500 (3)	0.4000 (6)
20	0.2000 (4)	0.3500 (7)

The actual number of relevant documents retrieved in n documents is shown in parentheses.

the original query for all n. The practical implication is that the augmented query is finding more relevant documents for the same number of retrieved documents. The actual number of relevant documents found at each n is shown in parentheses in the table. For example, the augmented query finds the same number of relevant documents (4) when retrieving ten documents as the original query finds when retrieving twenty documents. The augmented query finds three more relevant documents after 20 documents have been retrieved. This is tangible evidence of a potential effect. We quantify this effect over a larger set of queries in the remainder of the study.

6. Experimental methodology

In these experiments, we consider the effect of augmented queries on both document retrieval and collection selection. We also consider two paradigms for augmenting original queries. As a result, there are many experimental parameters. We begin with an overview of the three types of experiments, then cover the details of the experimental parameters.

6.1. Overview

6.1.1. Collection selection experiments. For the collection selection experiments, we evaluate collection selection independently of the eventual document retrieval at the selected collections. For these experiments, we are concerned with how augmented queries can affect our ability to locate collections that contain relevant documents. To study this, the primary experimental variable is the query formulation. We use the original query, then augment it with increasing numbers of MeSH terms and evaluate the results.

6.1.2. Document retrieval experiments. Our first document retrieval experiments mirror the collection selection experiments discussed above. Here, we are concerned with the effect of augmented queries on document retrieval. For the first experiments, we do not yet consider collection selection. Again, the primary experimental variable is the query formulation. We study document retrieval using the original query plus the original query augmented with MeSH terms when documents from *all collections* are eligible for retrieval.

6.1.3. Collection selection and document retrieval. The remaining experiments become more complicated and have more experimental variables. For these experiments, we consider the effects of augmented queries on document retrieval *when collection selection is also employed.* As a result, the queries used for both collection selection and document retrieval may vary. In addition, we use two different collection selection approaches.

6.2. Queries

We employ three different overall query formulations in these experiments. The first is the simple *original queries*, the statements of information need that are distributed with OHSUMED. The second formulation considers the original queries augmented with one, two or three top-ranked terms suggested by the *RBR-EVI* described above. The third type of query formulation is intended to simulate human-system interaction with an operational EVI. This approach was described in Section 5.3 and adds at most three MeSH terms to the original query.

For different experiments, we use different combinations of these approaches. For example, a query might be augmented for collection selection but the original query could be used for document retrieval.

6.3. Collection selection methodology

6.3.1. CORI. We used two collection selection approaches in our experiments. First, we used the existing *CORI* (Callan et al. 1995) collection selection approach. *CORI* has been shown to perform well for collection selection when compared to other approaches (Callan et al. 2000, French et al. 1999). *CORI* makes use of document frequency information about terms in collections to rank collections for selection. Because collection selection experiments were performed independently of document retrieval, we implemented the published *CORI* algorithm (Callan et al. 1995). The standard distribution of *CORI* operates in conjunction with the Inquery information retrieval system.

Given a set of databases to search, the *CORI* approach creates a *database selection index* in which each database is represented by its terms and their document frequencies *df*. Databases are ranked for a query q by a variant of the Inquery document ranking algorithm. The belief $p(r_k \mid c_i)$ in collection c_i due to observing query term r_k is determined by:

$$T = \frac{df}{df + 50 + 150 \cdot cw/\overline{cw}} \tag{1}$$

$$I = \frac{\log\left(\frac{|C|+0.5}{cf}\right)}{\log(|C| + 1.0)} \tag{2}$$

$$p(r_k \mid c_i) = 0.4 + 0.6 \cdot T \cdot I \tag{3}$$

where

df is the number of documents in c_i containing r_k,
cf is the number of collections containing r_k,

$|C|$ is the number of collections being ranked,
cw is the number of words in c_i, and
$\overline{cw}$ is the mean cw of the collections being ranked.

The belief in a database depends upon the query structure, but is usually just the average of the $p(r_k \mid c_i)$ values for each query term (Callan et al. 1995).

6.3.2. RBR. The second approach that we used was a relevance-based ranking (*RBR*) (French et al. 1998). This ranking served as an oracle for collection selection. Given the existence of relevance judgements, *RBR* ranks collections in descending order of the number of relevant documents that they contain. *RBR* is based upon the premise that it is advantageous to send queries to the collections containing the most relevant documents. It has been shown that multi-collection document retrieval improves markedly when *RBR* is used for collection selection (Craswell et al. 2000, Powell et al. 2000). One important thing to note is that because the ranking is based only upon the number of relevant documents in a collection, the *RBR* collection ranking for a query does not change if the query is augmented.

6.4. Document ranking

The document ranking formula used in all of these OHSUMED retrieval runs was the UC Berkeley TREC-2 probabilistic retrieval formula (Cooper et al. 1994). Retrieval results on the TREC test collections have shown that the formula is robust for both long queries and manually reformulated queries. The same formula (trained on English TREC collections) has performed well in other languages (Gey et al. 1996, Gey and Chen 1998, Gey et al. 1999, Chen et al. 1999). The algorithm has demonstrated its robustness independent of language as long as appropriate word boundary detection (segmentation) can be achieved. The logodds of relevance of document D to query Q is given by

$$\begin{aligned}\log O(R \mid D, Q) &= \log \frac{P(R \mid D, Q)}{P(\bar{R} \mid D, Q)}\\ &= -3.51 + \frac{1}{\sqrt{N}+1}\Phi + .0929 * N\end{aligned}$$

where

$$\begin{aligned}\Phi = 37.4 \sum_{i=1}^{N} \frac{qtf_i}{ql+35} + 0.330 \sum_{i=1}^{N} \log \frac{dtf_i}{dl+80}\\ - 0.1937 \sum_{i=1}^{N} \log \frac{ctf_i}{cl}\end{aligned}$$

where N is the number of terms overlapping between the query and document and qtf_i, dtf_i, ctf_i, ql, dl, and cl are term frequency in query, term frequency in document, collection term frequency for the ith matching term, and query length, document length, and collection length respectively. $P(R \mid D, Q)$ is the probability of relevance of document D with respect to query Q, $P(\bar{R} \mid D, Q)$ is the probability of irrelevance of document D with respect to query Q. Details about the derivation of these formulae may be found elsewhere (Cooper et al. 1994, Gey et al. 1996, Gey and Chen 1998, Gey et al. 1999, Chen et al. 1999).

6.5. Merging

In a multi-collection environment, collection selection is used to route queries to search engines at the individual collections. Merging the results from each collection into a single results list is an important, and often complex, problem. We avoid the difficulty of merging in our experimental environment by performing collection selection as a post-processing step following document retrieval. Specifically, we maintain all the documents in a centralized collection where each document is tagged with the collection to which it belongs. Documents are then retrieved from the centralized index. Documents from the selected collections are declared eligible for retrieval and the single results list is filtered to remove documents from other collections. In this case, no merge step is necessary. This approach is equivalent to a raw-score merge in a multi-collection environment where collection-wide information is available. See Powell et al. (2000) for a more detailed discussion of this approach.

6.6. Evaluation

As we discussed in Section 6.1, in this paper we report three types of experiments examining the effects of adding controlled vocabulary terms to queries. We examine the effects on collection selection performance, on document retrieval performance and on document retrieval when collection selection is employed. To study these effects, we utilize two different types of evaluation measures. When we focus on collection selection, we employ specialized collection selection performance measures that allow us to evaluate collection selection performance directly and independently of document retrieval performance. When we focus on document retrieval performance, we use traditional document retrieval evaluation techniques.

6.6.1. Collection selection. Our evaluation of collection selection approaches is based on the degree to which a collection ranking produced by an approach can approximate a desired collection ranking. Collection selection evaluation measures are discussed in detail in French and Powell (2000). For these experiments, we use only the $\mathcal{R}_n$ measure defined by Gravano and García-Molina (1995).

The $\mathcal{R}_n$ measure is calculated with respect to two rankings, a baseline ranking B that represents the desired collection ranking and an estimated ranking E produced by the collection selection approach. Our goal is to determine how well E approximates B. We assume that each collection C_i has some merit, *merit*(q, C_i), to the query q. The baseline is expressed in terms of this merit; the estimate is formed by implicitly or explicitly estimating merit. For these experiments, we always use a relevance-based ranking as the baseline, so *merit*(q, C_i) is the number of documents in C_i that are relevant with respect to query q.

Let C_{b_i} and C_{e_i} denote the collection in the ith ranked position of rankings B and E respectively. Let

$$B_i = merit(q, C_{b_i}) \quad \text{and } E_i = merit(q, C_{e_i}) \tag{4}$$

denote the merit associated with the ith ranked collection in the baseline and estimated rankings respectively.

Gravano and García-Molina (1995) defined $\mathcal{R}_n$ as follows.

$$\mathcal{R}_n = \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} B_i}. \tag{5}$$

This is a measure of how much of the available merit in the top n ranked collections of the baseline has been accumulated via the top n collections in the estimated ranking.

6.6.2. Document retrieval. For the document retrieval experiments reported here we use an approach that has been used for reporting TREC experimental results. We report precision at fixed numbers of documents retrieved. Precision is the number of relevant documents retrieved divided by the number of documents retrieved.

6.7. Summary of system configuration choices

We can now summarize our choices for the system configuration parameters identified in figure 1. We list these aspects below together with our choices.

1. *The query augmentation strategy QA.* We use *RBR-EVI* (Section 5.2) and the simulated user interaction SI (Section 5.3) to select MeSH terms for query augmentation. *RBR-EVI* is intended to reveal the maximum potential of the query augmentation approach while SI intends to show what performance effects would be seen by a knowledgeable user.
2. *The collection selection strategy CS.* We use *RBR* (Section 6.3.2) and *CORI* (Section 6.3.1) for collection selection algorithms. *RBR* is intended to demonstrate what is possible when an oracle is used for selection while *CORI* is illustrative of what is achievable with today's technology.
3. *The search engine technology.* We use the UC Berkeley TREC-2 probabilistic retrieval formula (Section 6.4) to rank documents for retrieval.

7. Results

We restate our hypotheses here and discuss the outcome of our experiments. In all the plots shown here, *RBR-EVI* is used to determine the "best" MeSH headings to use for query expansion. Results for a simulated user interaction (SI) are also reported. Where appropriate we show plots for both testbeds: the leftmost plot is always the 263 collection testbed decomposed by journal; the rightmost plot is the 48 collection testbed decomposed by subject category.

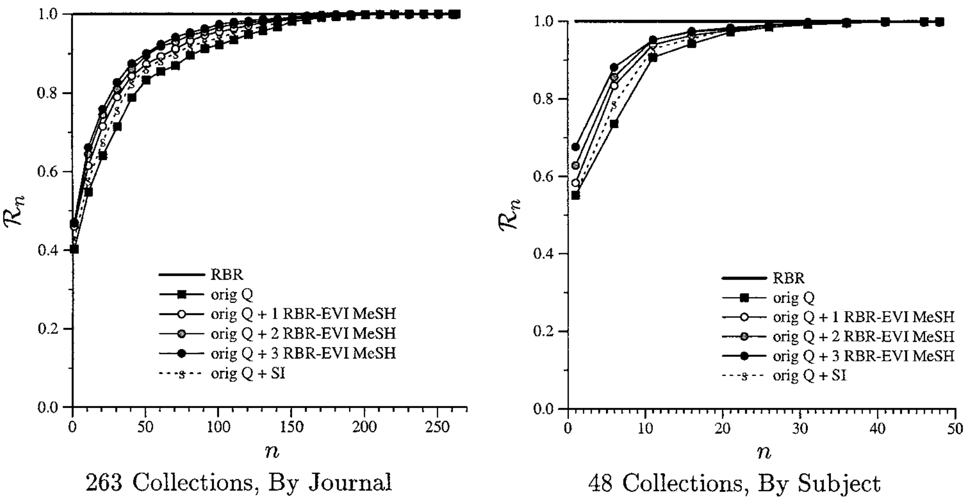


Figure 5. *CORI* selection performance measured by $\mathcal{R}_n$ when 0, 1, 2 or 3 MeSH terms are added to the original query for collection selection.

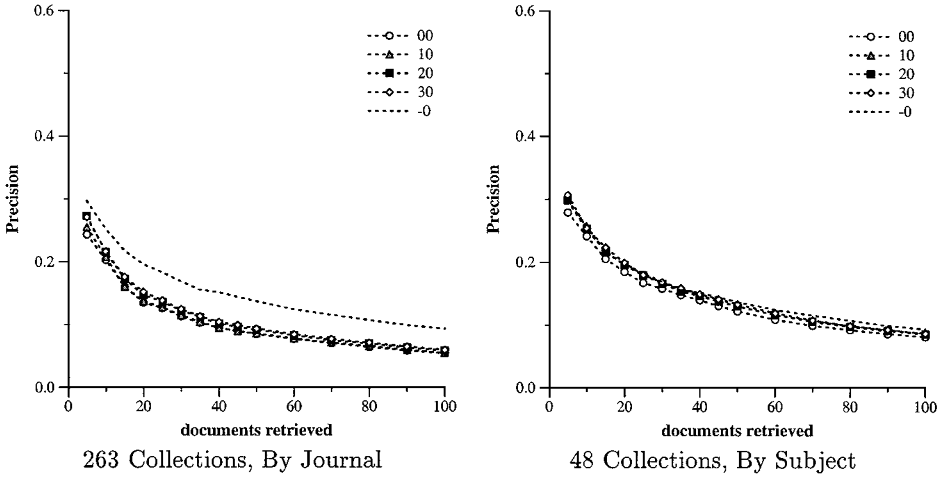
7.1. Collection selection

Hypothesis 1. Augmented queries will be more effective for collection selection than the original queries. Adding more MeSH headings will improve selection results.

For these experiments, we evaluated directly the effect of augmenting queries on collection selection. Here, we used the $\mathcal{R}_n$ measure for evaluation; no document retrieval has been performed yet. Figure 5 shows the results of our collection selection comparison and illustrates three different types of queries: the original queries, the original queries augmented by *RBR-EVI* MeSH terms and the original queries augmented using SI.

We used the *CORI* algorithm (Callan et al. 1995) to perform collection selection because prior research has shown it to be as good as or superior to other collection selection algorithms (Powell et al. 2000, French et al. 1998, 1999). For contrast, the best possible performance under the $\mathcal{R}_n$ measure is shown as the curve labeled *RBR*. Note that $\mathcal{R}_n = 1$ for *RBR*.

As can clearly be seen from figure 5, *Hypothesis 1* is born out. When the *RBR-EVI* is used to augment queries, the addition of MeSH terms to the original query boosts collection selection performance by over 25% up to about 70 document collections selected for the 263 collection testbed. The improvement beyond that is somewhat smaller but still significant. Adding more *RBR-EVI* MeSH terms does improve collection selection performance but the magnitude of improvement drops off after two terms have been added. A visible improvement can also be observed when the simulated interaction (SI) approach to query augmentation is employed, suggesting that a portion of the potential improvement shown under *RBR-EVI* is achievable in an operational setting.



263 Collections, By Journal

48 Collections, By Subject

Figure 6. Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used for collection selection but the original query is used for document retrieval. This scenario corresponds to figure 1(d). Note that the line labeled "-0" corresponds to figure 1(a) while the line labeled "00" corresponds to figure 1(c).

> *Hypothesis 2.* The benefits of using augmented queries for collection selection will translate to superior document retrieval results, even when the original queries are used for document retrieval.

For these experiments, we used the *CORI* collection selection rankings whose performance was evaluated in figure 5 and selected the 5 top-ranked collections for each query. Retrieval was restricted to the documents contained in those collections. We varied the query formulation used for collection selection, but always used the original query for document retrieval. In an operational setting, it is likely that augmented queries would be used for document retrieval; however, in this case we wanted to isolate the effect of the augmented queries when used for collection selection.

As in the experiments reported for *Hypothesis 1*, we used the *RBR-EVI* to augment the queries with 1, 2 or 3 MeSH terms. Before we examine figure 6, it is necessary to explain the labeling convention of our figures. Each plot on the graphs of figures 6–9 is labeled according to the number of MeSH terms added to the original query. The first digit of the label is the number of terms added to the collection selection query. The second digit is the number of terms added to the document retrieval query. For example, plot "20" of figure 6 shows results when two *RBR-EVI* MeSH terms were added to the collection selection query and when zero MeSH terms were added to the document retrieval query (i.e. the original query was used). There are a few additions to this convention. We use "-" to denote no collection selection step and "*" to denote *RBR* selection (recall that *RBR* selection is not affected by query augmentation). An "s" denotes the use of SI augmented queries for either collection selection or document retrieval. The line and mark types of the plots are also consistent across figures 6–9. Please note that the Precision values in figures 6–9 have maximum value of 0.6 to facilitate graph readability.

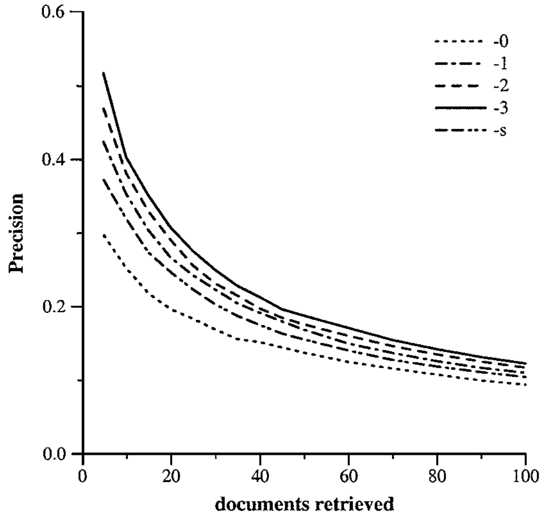


Figure 7. Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used to augment the query for retrieval. No collection selection used. This scenario corresponds to figure 1(b). Note that the line labeled “-0” corresponds to figure 1(a).

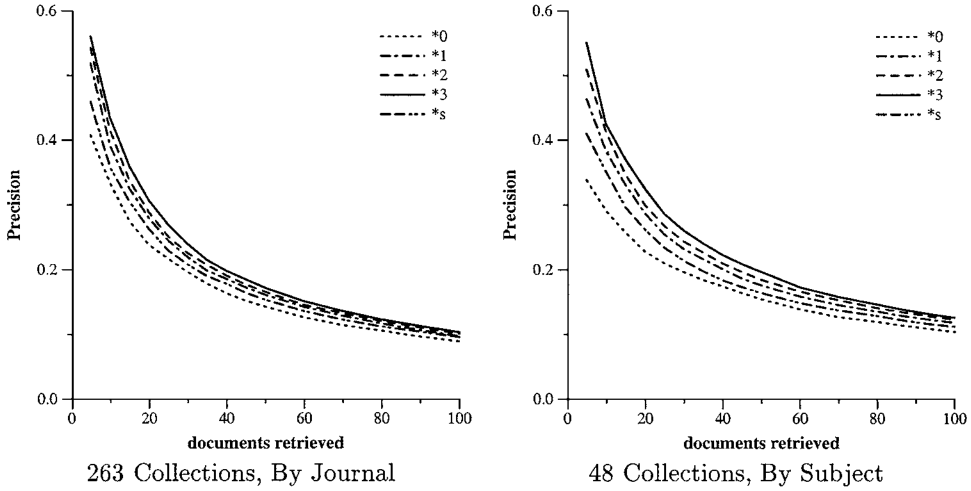
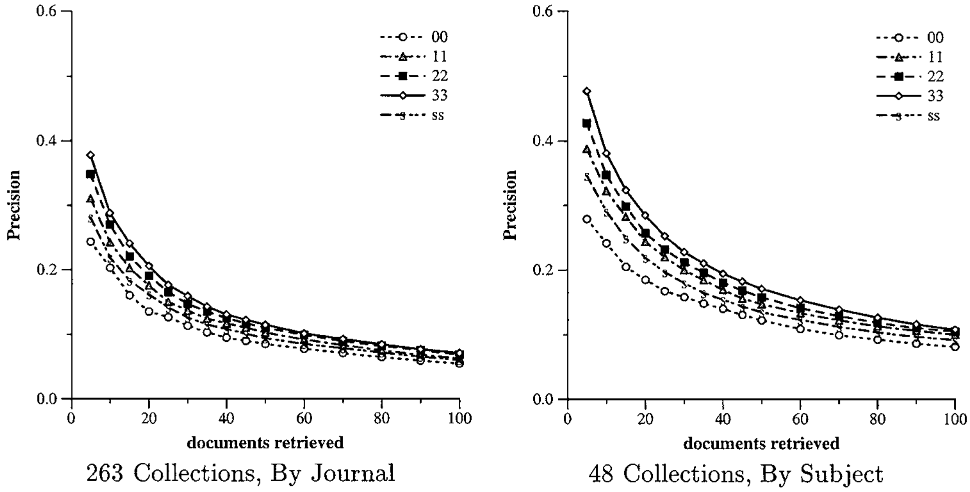


Figure 8. Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used for for document retrieval and an oracle is used for collection selection. This scenario corresponds to figure 1(f) where CS is achieved by *RBR*.

We see from figure 6 that *Hypothesis 2* is false. Consider first the journal decomposition. While there is some slight improvement in retrieval performance as MeSH terms are added for collection selection, the performance overall is largely unchanged from that using the original query alone. For example, when the original query is used for both collection



263 Collections, By Journal

48 Collections, By Subject

Figure 9. Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used for collection selection and for document retrieval. This scenario corresponds to figure 1(e). Note that the line labeled "00" corresponds to figure 1(c).

selection and document retrieval, precision at 20 documents retrieved is 0.13 for the 263 collection testbed. Adding 1, 2 and 3 MeSH terms yields precision values of 0.14, 0.15 and 0.15 respectively. For comparison we have also shown the performance obtained when the original query is used on the unpartitioned document collection (the plot labeled "-0"). Note that the collection selection approach is searching $5/263 < 2\%$ of the document collections and while its performance is lower than that obtained by searching the unpartitioned collection, it is still quite respectable given the small number of collections searched.

The conclusion is the same with respect to the subject category decomposition but the situation is much brighter here. While adding MeSH terms did not appreciably improve the performance of collection selection over the original query, collection selection did perform almost as well as the original query on the unpartitioned data. However in this case collection selection confined the search to approximately 10% (5/48) of the collections. The clear indication here is that we can search subsets of documents with similar retrieval performance to searches having all the documents at their disposal.

7.2. Document retrieval

> *Hypothesis 3.* Augmented queries will outperform the original queries for document retrieval.

For these experiments, no collection selection step was performed. All documents were eligible for retrieval.

Inspecting figure 7 we see that *Hypothesis 3* is clearly correct. The single "best" MeSH term suggested by the *RBR-EVI* caused a large performance boost with smaller gains coming from the addition of more terms. The achievable SI approach (marked with "s" in the plot) fell short of the oracle, but achieved a significant performance boost over the original query. We conclude that a user familiar with the controlled vocabulary would benefit from the term suggestions of an EVI.

8. Discussion

Collection selection and document retrieval are two different problems and techniques improving one need not improve the other. This is seen clearly in figures 6 and 7. Adding MeSH terms to the query for collection selection alone had little effect on final document retrieval performance (figure 6); augmenting a query with MeSH terms for document retrieval showed substantial performance gains (figure 7).

We ran another experiment to get an idea of what kind of performance gain is possible using an oracle for collection selection and using augmented queries for document retrieval. In this experiment collection selection was determined by the *RBR* approach (French et al. 1998). The results are shown in figure 8. A comparison with figure 7 shows that additional performance gains are possible when excellent collection selection is employed.

The obvious question to ask now is: what kind of performance is achievable using existing collection selection technology and augmented queries. Figure 9 shows the retrieval performance when *CORI* selection is used together with both *RBR-EVI* and SI augmented queries for both collection selection and document retrieval. Recall that the results when collection selection is employed are computed over less than 2% of the 263 document collections and over approximately 10% of the 48 collections so figure 9 should be compared to figure 6. We see that the addition of more MeSH terms improves retrieval performance. Moreover, this strategy is comparable to the performance of the original query on the unpartitioned collection through approximately 40 documents retrieved. In these testbeds only 16 of the 101 queries have as many as 40 relevant documents.

We also note from figure 7 that the best performance of augmented queries (i.e., when 3 MeSH terms are added) is everywhere better than the best performance shown in figure 9, but we emphasize again, only a small percentage of the document collections are used for retrieval by the strategy employed in figure 9. The results in figure 9 reinforce earlier work demonstrating that good retrieval performance can be obtained even when the search space is severely restricted (Powell et al. 2000).

9. Conclusions and future research

Our paper has contributed to the understanding of query augmentation in collection selection and document retrieval.

This research has addressed the question of exploitation of controlled vocabulary to improve information retrieval performance from both a distributed collection selection and document retrieval point of view. We have shown that intelligent query expansion by augmenting natural language queries with controlled vocabulary terms can result in significant performance improvement (demonstrated by the results in figure 7). The augmentation is achieved in practice through the use of Entry Vocabulary Indexes (EVIs) which map from ordinary language expressions to controlled vocabulary index terms. When index term suggestions are reviewed interactively by a human, the most effective terms can be selected from many presented by the EVI system.

An evaluation methodology has been presented which simulates human selection by its overlap between relevance-based *RBR-EVI* performance and actual ranked lists of EVI suggested terms for query expansion. The assumptions underlying this strategy are reasonable and when it can be used, the strategy gives us a means to deterministically evaluate interactive retrieval performance. We have shown that the simulated interactive query expansion from a controlled vocabulary can gain as much as 30 percent over the original free text query. The results, of course, apply to document collections which possess the value-added augmentation of human indexing. This, however, covers much of the existing scientific literature and hence techniques which improve technical literature search are intrinsically worthwhile.

As shown in figure 8, collection selection has the potential to radically increase document retrieval performance. Today's technology is only achieving a small portion of that potential. Research into better collection selection algorithms is clearly worthwhile.

An open research question is whether a methodology for automatic query expansion can be found which achieves some of the value-added of human term selection for expansion. If, for example, the subdomain of discourse (say, for example, *Surgery*, with respect to the medical literature) could be identified, the controlled vocabulary could be restricted to that subdomain, and further performance improvements might be attainable. This is one direction of our current research.

Appendix A: Data preprocessing

We had to deal with a number of syntactic issues to ensure proper handling of the data. MeSH categories from different levels of the hierarchy are used in OHSUMED as individual instances. For example, `Accidents` and `Accidents, Occupational` are both used even though the latter is a subcategory of the first. The comma does not denote subcategory. For example, `Drowning` is also a subcategory of `Accidents`, but does not appear as `Accidents, Drowning`.

The MeSH terms often have a two character topical subheading qualifier[3] attached. These codes appear following slashes in the category name. For example `MO` stands for mortality and `Accidents, Traffic/MO` is used to describe fatal traffic accidents. In some cases an asterisk is used to denote the principal MeSH heading assigned to the article, e.g., `Accidents, Traffic/*MO`. Table 2 shows some additional examples of these qualified MeSH terms together with the expansions of the qualifiers.

When parsing the data we replaced spaces, periods, slashes, dashes, apostrophes and & with underscores and deleted commas, plus signs, trailing periods, and brackets. Since slashes are replaced by underscores, all qualifiers are preceded by underscores.

Table 4 shows an example of the way in which we preprocessed the data. The leftmost column represents the maximal set of terms that we could have dealt with while the rightmost column is the set we actually chose to use in the experiments.

Table 4. The leftmost column represents the maximal set of preprocessed terms. The rightmost column is the set of principal MeSH terms indicated by "*" in the original text.

Original text of MeSH terms assigned to article	
Adult; Case Report; Cauda Equina/*; Hemangioma, Cavernous/*CO/PA; Human; Male; Myelography; Nuclear Magnetic Resonance/DU; Peripheral Nerve Neoplasms/*CO/PA; Subarachnoid Hemorrhage/*ET; Tomography, X-Ray Computed; Non-U.S. Gov't.	
All MeSH terms assigned	Principal terms assigned
adult	
case_report	
cauda_equina	cauda_equina
hemangioma_cavernous_co	hemangioma_cavernous_co
hemangioma_cavernous_pa	
human	
male	
myelography	
nuclear_magnetic_resonance_du	
peripheral_nerve_neoplasms_co	peripheral_nerve_neoplasms_co
peripheral_nerve_neoplasms_pa	
subarachnoid_hemorrhage_et	subarachnoid_hemorrhage_et
tomography_x_ray_computed	
non_us_govt	

Only the most applicable MeSH headings were retained for use in the representations for collection selection and document retrieval. These are indicated by an asterisk.

Appendix B: OHSUMED 48 collection decomposition

This appendix gives the explicit assignment of journals to subject areas that we used for the 48 collection decomposition of the OHSUMED testbed.

Alcoholism
Alcohol Alcohol
Alcohol Clin Exp Res
J Stud Alcohol
J Subst Abuse Treat
Am J Hum Genet
Am J Hum Genet
Anesthesiology
Anaesthesia
Anesth Analg
Anesthesiology
Br J Anaesth
Can J Anaesth
Int Anesthesiol Clin
Arthritis
Arthritis Rheum
Br J Rheumatol
J Rheumatol
Artificial Organs
ASAIO Trans

Int J Artif Organs
Transplant Proc

Burns
Burns Incl Therm Inj
J Burn Care Rehabil

Cardiology
Am Heart J
Am J Cardiol
Br Heart J
Circulation
Curr Probl Cardiol
J Am Coll Cardiol
J Am Soc Echocardiorgr

Cardiovascular Diseases
Cardiovasc Clin
Cathet Cardiovasc Diagn
Circ Res
J Cardiovasc Surg Torino
J Thorac Cardiovasc Surg
J Vasc Surg
Prog Cardiovasc Dis
Scand J Thorac Cardiovasc Surg

Communicable Diseases
AIDS
AIDS Res Hum Retroviruses
Am J Trop Med Hyg
Antimicrob Agents Chemother
Infect Dis Clin North Am
J Acquir Immune Defic Syndr
J Infect Dis
Rev Infect Dis

Dentistry
J Am Dent Assoc
J Oral Maxillofac Surg
Oral Surg Oral Med Oral Pathol

Dermatology
Arch Dermatol
Br J Dermatol
Contact Dermatitis
Curr Probl Dermatol
J Am Acad Dermatol
J Dermatol Surg Oncol
J Invest Dermatol

Diabetes Mellitus
Diabetes
Diabetes Care

Drug Therapy
Clin Pharmacol Ther
DICP
Drug Intell Clin Pharm
J Pharmacol Exp Ther
J Psychoactive Drugs
Med Lett Drugs Ther
Pharmacol Rev

Education Medical
Acad Med
J Med Educ

Emergency Medicine
Am J Emerg Med
Ann Emerg Med
Emerg Med Clin North Am
J Emerg Med
J Toxicol Clin Toxicol
J Trauma

Endocrinology
Endocrinology
J Clin Endocrinol Metab

Family Practice
Am Fam Physician
Fam Med
Fam Pract
Fam Pract Res J
J Am Board Fam Pract
J Fam Pract

Gastroenterology
Am J Gastroenterol
Dig Dis Sci
Gastroenterology
Gastrointest Endosc
Gut
Hepatology
J Clin Gastroenterol

Geriatrics
Geriatrics
Gerontologist
J Am Geriatr Soc
J Gerontol

Gynecology Obstetrics
Am J Obstet Gynecol

Br J Obstet Gynaecol
Clin Obstet Gynecol
Clin Perinatol
Fertil Steril
J In Vitro Fert Embryo Transf
J Reprod Med
Obstet Gynecol
Obstet Gynecol Clin North Am
Surg Gynecol Obstet

Hematology
Blood
Transfusion

Hypersensitivity
Ann Allergy
Clin Rev Allergy
J Allergy Clin Immunol
J Immunol

Hypertension
Am J Hypertens
Hypertension

Intensive Care Units
Crit Care Med
Heart Lung
J Clin Monit

Internal Medicine
Acta Med Scand
Adv Intern Med
Am J Med
Am J Med Sci
Ann Intern Med
Arch Intern Med
BMJ
Br Med J Clin Res Ed
Can Med Assoc J
Clin Sci
Dis Mon
J Clin Invest
J Gen Intern Med
J Intern Med
JAMA
Lancet
Mayo Clin Proc
Med Clin North Am
Medicine Baltimore
N Engl J Med
Postgrad Med
Prim Care
Q J Med
South Med J
West J Med

J Appl Physiol
J Appl Physiol

Kidney Diseases
Am J Kidney Dis
Clin Nephrol
J Am Soc Nephrol
Kidney Int
Kidney Int Suppl

Laboratory Techniques and Procedures
Am J Clin Pathol
Am J Forensic Med Pathol
Am J Pathol
Arch Pathol Lab Med
Clin Lab Med
J Clin Pathol
J Forensic Sci
J Lab Clin Med
J Neuropathol Exp Neurol

Lasers–therapeutic use
Lasers Surg Med
Lasers Surg Med Suppl

MMWR Morb Mortal Wkly Rep
MMWR Morb Mortal Wkly Rep

Neoplasms
CA Cancer J Clin
Cancer
J Clin Oncol
J Natl Cancer Inst
J Surg Oncol

Neurology
Ann Neurol
Arch Neurol
Brain
Dysphagia
Epilepsia
Headache
J Neurol
Muscle Nerve

Neurol Clin
Neurology
Pain
Spine
Stroke

Neurosurgery

J Neurol Neurosurg Psychiatry
J Neurosurg
Neurosurgery
Surg Neurol

Nursing

Am J Nurs
J Nurs Adm
MCN Am J Matern Child Nurs
Nurs Clin North Am
Nurs Outlook
Nurs Res

Nutrition

Am J Clin Nutr
J Am Diet Assoc
J Nutr
JPEN J Parenter Enternal Nutr
Nutr Clin Pract
Nutr Rev

Ophthalmology

Am J Ophthalmol
Ann Ophthalmol
Arch Ophthalmol
Br J Ophthalmol
Can J Ophthalmol
Ophthalmic Surg
Ophthalmologica
Ophthalmology
Surv Ophthalmol

Orthopedics

Clin Orthop
J Bone Joint Surg Am
J Bone Joint Surg Br
Orthop Clin North Am

Otolaryngology

Ann Otol Rhinol Laryngol
Arch Otolaryngol Head Neck Surg
Clin Otolaryngol
Head Neck Surg
J Laryngol Otol
Laryngoscope
Otolaryngol Clin North Am
Otolaryngol Head Neck Surg

Pediatrics

Adv Pediatr
Am J Dis Child
Arch Dis Child
Clin Pediatr Phila
J Pediatr
Pediatr Clin North Am
Pediatr Emerg Care
Pediatr Infect Dis J
Pediatr Neurol
Pediatrics

Physical Therapy

Am J Phys Med
Am J Phys Med Rehabil
Arch Phys Med Rehabil
Compr Ther
Phys Ther

Psychiatry

Am J Psychiatry
Arch Gen Psychiatry
J Nerv Ment Dis

Public Health

Am J Public Health
Arch Environ Health
Health Care Manage Rev
Hospitals
J Clin Epidemiol
MD Comput
Public Health Rep

Radiology

AJR Am J Roentgenol
Angiology
Br J Radiol
J Nucl Med
Radiol Clin North Am
Radiology

Science

Nature
Proc Natl Acad Sci USA
Sci Am

Science

Sports Medicine

Am J Sports Med
Clin Sports Med
Med Sci Sports Exerc

Surgery

Am J Surg
Am Surg
Ann Surg
Arch Surg
Br J Surg
Curr Probl Surg
Dis Colon Rectum
Plast Reconstr Surg
Surg Clin North Am
Surgery

Thoracic Diseases

Am Rev Respir Dis
Ann Thorac Surg
Chest
Thorax

Urology

Br J Urol
J Urol
Urol Clin North Am
Urology

Acknowledgments

We thank Ray Larson and Aitao Chen of UC Berkeley for helpful observations on the design of this EVI evaluation. We would also like to thank William Hersh and Paul Gorman of Oregon Health and Science University for their comments and help with the decomposition of the OHSUMED data shown in Appendix B.

Notes

1. The principal MeSH terms are those that were denoted by the indexer as being central to the article. We discuss our treatment of these terms in more detail in Appendix A.
2. Magennis and van Rijsbergen (1997) showed that for non-controlled vocabulary the full benefit may not be achieved by inexperienced users.
3. A list of the topical subheadings is available at `http://www.nlm.nih.gov/mesh/topcat.html`. Some of the topical subheadings employ abbreviations that are described at `http://www.nlm.nih.gov/mesh/abbrev2002.html`.

References

Buckland M et al. (1999) Mapping entry vocabulary to unfamiliar metadata vocabularies. In: *D-Lib Magazine*. http://www.dlib.org/dlib/january99/buckland/01buckland.html.

Callan J, Powell AL, French JC and Connell M (2000) The effects of query-based sampling on automatic database selection algorithms. Technical Report CMU-LTI-00-162, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Callan JP, Lu Z and Croft WB (1995) Searching distributed collections with inference networks. In: Proc. ACM SIGIR'95, pp. 21–28.

Callan JP, Lu Z and Croft WB (1995) Searching distributed collections with inference networks. In: Proc. SIGIR'95, pp. 21–29.

Chen A, Kishida K, Jiang H, Liang Q and Gey FC (1999) Comparing multiple methods for Japanese and Japanese-English text retrieval. In: First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp. 49–58.

Cooper WS, Chen A and Gey FC (1994) Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In: Text REtrieval Conference (TREC-2), pp. 57–66.

Craswell N, Bailey P and Hawking D (2000) Server selection on the world wide web. In: Proc. ACM Digital Libraries Conf., pp. 37–46.

Doszkocs TE (1983) CITE NLM: Natural language searching in an online catalog. Information Technology and Libraries, 2:364–380.

Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74.

Eichmann D, Ruiz M and Srinivasan P (1998) Cross-language information retrieval with the UMLS metathesaurus. In: Proc. ACM SIGIR'98, pp. 72–80.

French JC and Powell AL (2000) Metrics for evaluating database selection techniques. World Wide Web: Internet and Web Information Systems, 3(3).

French JC, Powell AL, Callan J, Viles CL, Emmitt T, Prey KJ and Mou Y (1999) Comparing the performance of database selection algorithms. In: Proc. ACM SIGIR'99, pp. 238–245.

French JC, Powell AL, Gey F and Perelman N (2001) Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness. In: Proc. Tenth International Conference on Information and Knowledge Management (CIKM 2001), pp. 199–206.

French JC, Powell AL, Viles CL, Emmitt T and Prey KJ (1998) Evaluating database selection techniques: A testbed and experiment. In: Proc. ACM SIGIR'98, pp. 121–129.

Fuhr N (1999) A decision-theoretic approach to database selection in networked IR. ACM Transactions on Information Systems, 17(3):229–249.

Gauch S, Wang J and Rachakonda SM (1999) A corpus analysis approach for automatic query expansion and its extension to multiple databases. ACM Transactions on Information Systems, 17(3):250–269.

Gey F, Buckland M, Chen A and Larson R (2001) Entry vocabulary—A technology to enhance digital object search. In: Proceedings of the First Internation Conference on Human Language Technology.

Gey F, Jiang H, Chen A and Larson R (1999) Manual queries and machine translation in cross language retrieval and interactive retrieval at TREC-7. In: Text REtrieval Conference (TREC-7), pp. 527–539.

Gey FC and Chen A (1998) Phrase discovery for English and cross-language retrieval at TREC-6. In: Text REtrieval Conference (TREC-6), pp. 637–648.

Gey FC, Chen A, He J, Xu L and Meggs J (1996) Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: Probabilistic algorithms at TREC-5. In: Text Retrieval Conference (TREC-5).

Gravano L and García-Molina H (1995) Generalizing GlOSS to vector-space databases and broker hierarchies. In: Proc. of the 21st VLDB Conference, pp. 78–89.

Gravano L, García-Molina H and Tomasic A (1999) GlOSS: Text-source discovery over the internet. ACM Trans. on Database Systems, 24(2):229–264.

Harman D (1988) Towards interactive query expansion. In: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 321–331.

Hawking D and Thistlewaite P (1999) Methods for information server selection. ACM Transactions on Information Systems, 17(1):40–76.

Hersh W, Buckley C, Leone TJ and Hickam D (1994) OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Proc. ACM SIGIR'94, pp. 192–201.

Hersh W, Price S and Donohoe L (2000) Assessing thesaurus-based query expansion using the UMLS metathesaurus. In: Proceedings of the 2000 American Medical Informatics Association (AMIA) Symposium.

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H and Hidaka S (1999) Overview of IR tasks at the first NTCIR workshop. In: The First NTCIR Workshop on Japanese Text Retrieval and Term Recognition, pp. 11–22.

Kluck M and Gey F (2001) The domain-specific task of CLEF—Specific evaluation strategies in cross-language information retrieval. In: Cross-Language Information Retrieval Evaluation, Proceedings of the CLEF 2000 Workshop, Forthcoming. Springer.

Magennis M and van Rijsbergen CJ (1997) The potential and actual effectiveness of interactive query expansion. In: SIGIR'97, pp. 324–332.

Meng W, Liu K-L, Yu C, Wang X, Chang Y and Rishe N (1998) Determining text databases to search in the internet. In: Proceedings of the 24th VLDB Conference, pp. 14–25.

Powell AL, French JC, Callan J, Connell M and Viles CL (2000) The impact of database selection on distributed searching. In: Proc. ACM SIGIR '00, pp. 232–239.

Schatz B, Chen H et al. (1996) Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurence lists for information retrieval. In: Proc. ACM Digital Libraries Conf.

Schott H (ed.) (2000) Thesaurus for the Social Sciences. (Vol. 1:) German-English. (Vol. 2:) English-German. (Edition) 1999. InformationsZentrum Sozialwissenschaften Bonn.

Xu J and Callan J (1998) Effective retrieval with distributed collections. In: Proc. ACM SIGIR'98, pp. 112–120.

Yu C, Meng W, Liu K-L, Wu W and Rishe N (1999) Efficient and effective metasearch for a large number of text databases. In: Proc. ACM CIKM'99, pp. 217–224.

Yuwono B and Lee DL (1997) Server ranking for distributed text retrieval systems on internet. In: Proceedings of the Fifth International Conference on Database Systems for Advanced Applications, pp. 41–49.