Exploiting a Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness

James C. French*
Dept. of Computer Science
University of Virginia
Charlottesville, VA
french@cs.virginia.edu

Allison L. Powell†
Corporation for National Research Initiatives
Reston, VA
apowell@cnri.reston.va.us

Fredric Gey Natalia Perelman‡
UC Data Archive & Technical Assistance
University of California
Berkeley, CA
gey@ucdata.berkeley.edu

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.

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*.

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 a multi-collection test environment based on the OHSUMED [22] test collection and discuss features of that test environment. 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 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. Background and Related Work

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

†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.

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.

2.1 Manual Indexing

The OHSUMED collection [22], 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. Other collections using controlled vocabulary include NTCIR [24] and GIRT (German Information Retrieval Test) [25]. 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 [7], 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 [29]. Gey *et al.* [14] 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) [1, 14].

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 [23]. 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 choose which words, phrases, or terms are to be added to the query.

2.2 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 [3, 12, 19, 21, 27, 31, 32]. Three of these approaches, *CORI*[3], *CVV*[32] and *gGlOSS*[19] were evaluated in a common environment by French, *et al.*[2, 10, 11], 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 [28, 30]. Of particular interest to us here is the work of Xu and Callan [30] who noted that query expansion can improve collection selection performance. Xu and Callan studied query expansion using the general vocabulary of document in the collections, however in this work we consider the effect of augmented queries using controlled vocabulary.

3. OHSUMED-based test environment

All of the experiments reported here were conducted using a specific organization of the documents found in the OHSUMED test collection. The OHSUMED collection, constructed and described by Hersh *et al.* [22], 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 assigned.

The manually-assigned MeSH headings make the OHSUMED collection useful for our study of augmented queries; however, we're 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 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).

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

3.1 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.

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.

3.2 Distribution of Documents and Relevant Documents

The OHSUMED test collection, and our journal-based organization of the documents into a multi-collection testbed 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.

Our choice of organizing the documents by publishing journal resulted in a skewed distribution of documents among collections. On average, there are 1,325 documents per collection with a minimum of 3 and a maximum of 12,654. 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. 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 [9].

4. 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. The second approach augments the queries with the MeSH terms most frequently assigned to relevant documents.

4.1 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 [13]. 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.

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

Table 1: Contingency table from words/phrases to classification

The association score between a word/phrase t and a classification C is computed following Dunning [8]

$$W(C,t) = 2[logL(p_1,a,a+b) + logL(p_2,c,c+d) - logL(p,a,a+b) - logL(p,c,c+d)]$$

where

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

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

4.2 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 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.

4.3 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[1]. 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 [20] 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 6.

5. 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.

5.1 Overview

5.1.1 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.

5.1.2 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 263 collections* are eligible for retrieval.

5.1.3 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.

5.2 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 4.3 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.

5.3 Collection Selection Methodology

We used two collection selection approaches in our experiments. First, we used the existing *CORI* [3] collection selection approach. *CORI* has been shown to perform well for collection selection when compared to other approaches [2, 10]. *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 [3]. The standard distribution of *CORI* operates in conjunction with the Inquery information retrieval system.

The second approach that we used was a relevance-based ranking (RBR). 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 [6, 28]. 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.

5.4 Document Ranking

The document ranking formula used in all of these OHSUMED retrieval runs was the UC Berkeley TREC-2 probabilistic retrieval formula [5]. 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 [17, 16, 15, 4]. The algorithm has demonstrated its robustness independent of lan-

[1]Magennis and van Rijsbergen [26] showed that for non-controlled vocabulary the full benefit may not be achieved by inexperienced users.

guage as long as appropriate word boundary detection (segmentation) can be achieved. The logodds of relevance of document D to query Q is given by

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

where

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

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. $P(R|D,Q)$ is the probability of relevance of document D with respect to query Q, $P(\overline{R}|D,Q)$ is the probability of irrelevance of document D with respect to query Q. Details about the derivation of these formulae may be found elsewhere [5, 17, 16, 15, 4].

5.5 Merging

There are two ways that collection selection can be employed. In an existing multi-collection environment, collection selection is used to route queries to search engines at the individual collections. In this case, merging the results from each collection into a single results list is an important, and often complex, problem.

When the documents from all collections are available (as is the case here), collection selection can be performed as a post-processing step when documents are retrieved from a centralized index of the documents in all collections. Documents from the selected collections can be declared eligible for retrieval and the single results list is filtered. In this case, no merge step is necessary. This is the approach employed for the experiments reported here. This approach is equivalent to a raw-score merge in a multi-collection environment where collection-wide information is available. See Powell *et al.* [28] for a more detailed discussion of this approach.

5.6 Evaluation

5.6.1 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 [9]. For these experiments, we use only the $\mathcal{R}_n$ measure defined by Gravano and García-Molina [18].

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 i-th ranked position of rankings B and E respectively. Let

$$B_i = merit(q, C_{b_i}) \text{ and } E_i = merit(q, C_{e_i}) \qquad (1)$$

denote the merit associated with the i-th ranked collection in the baseline and estimated rankings respectively.

Gravano *et al.*[18] defined $\mathcal{R}_n$ as follows.

$$\mathcal{R}_n = \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} B_i}. \qquad (2)$$

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.

5.6.2 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. 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.

6.1 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 1 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 [3] to perform collection selection because prior research has shown it to be as good as or superior to other collection selection algorithms[28, 10, 11]. For contrast, the best possible performance under the $\mathcal{R}_n$ measure is shown as the curve labeled RBR.

As can clearly be seen from Figure 1, 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. 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.

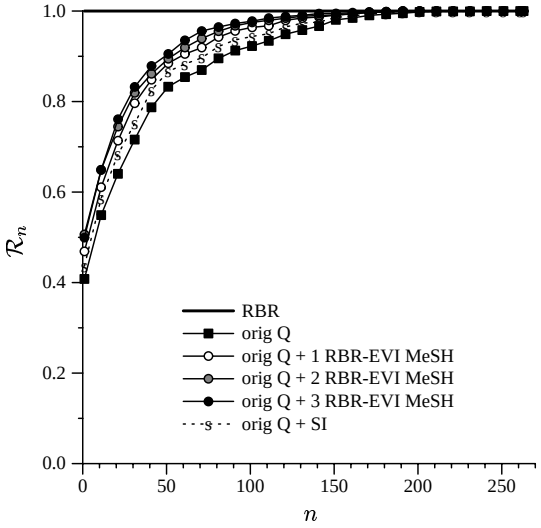

Figure 1: CORI selection performance measured by $\mathcal{R}_n$ when 0, 1, 2 or 3 MeSH terms are added to the original query for collection selection.

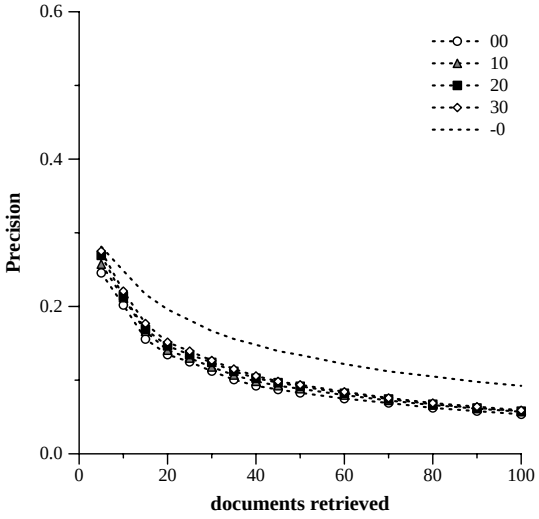

Figure 2: Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used for collection selection but the original query is used for document retrieval.

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 1 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 2, it is necessary to explain the labeling convention of our figures. Each plot on the graphs of Figures 2–5 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 2 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 2–5. Please note that the Precision values in Figures 2–5 have maximum value of 0.6 to facilitate graph readability.

We see from Figure 2 that this Hypothesis 2 is false. 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 selection and document retrieval, precision at 20 documents retrieved is 0.13. 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.

6.2 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 from the 263 collections were eligible for retrieval.

Inspecting Figure 3 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.

7. Discussion

Collection selection and document retrieval are two different problems and techniques improving one need not improve the other. This is seen clearly in Figure 2 and Figure 3. Adding MeSH terms to the query for collection selection alone had little effect on final document retrieval performance; augmenting a query with MeSH terms for document retrieval showed substantial performance gains.

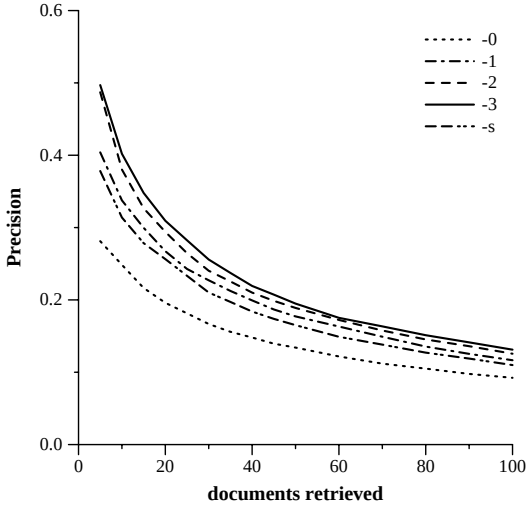

Figure 3: Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used to augment the query for retrieval. No collection selection used.

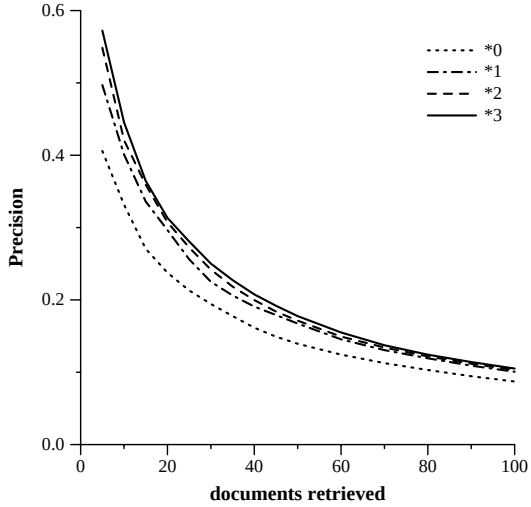

Figure 4: Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used for for document retrieval and an oracle is used for collection selection.

We ran another experiment to get an idea of what kind of performance gain is possible with the best possible collection selection and using augmented queries for document retrieval. In this experiment collection selection was determined by the RBR approach[11]. The results are shown in Figure 4. A comparison with Figure 3 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 today's best collection selection technology and augmented queries. Figure 5 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 so Figure 5 should be compared to Figure 2. 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 this testbed only 16 of the 101 queries have as many as 40 relevant documents.

We also note from Figure 3 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 5, but we emphasize again, only 2% of the document collections are used for retrieval by the strategy employed in Figure 5. The results in Figure 5 reinforce earlier work demonstrating that good retrieval performance can be obtained even when the search space is severely restricted[28].

8. 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 3). 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 4, 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.

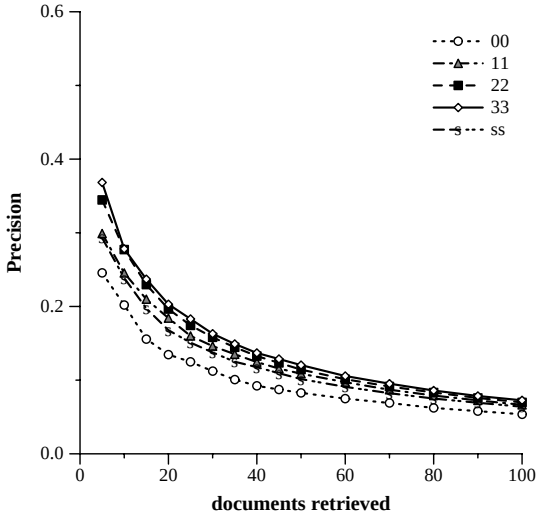

Figure 5: Document retrieval performance measured by average precision when 0, 1, 2 or 3 MeSH terms are used for collection selection and for document retrieval.

Acknowledgements

We thank Ray Larson and Aitao Chen of UC Berkeley for helpful observations on the design of this EVI evaluation.

References

[1] M. Buckland et al. Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. In *D-Lib agazine*, January 1999. http://www.dlib.org/.

[2] J. Callan, A. L. Powell, J. C. French, and M. Connell. 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, 2000.

[3] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In *Proc. AC SIGIR '95*, pages 21–28, 1995.

[4] A. Chen, K. Kishida, H. Jiang, Q. Liang, and F. C. Gey. Comparing multiple methods for Japanese and Japanese-English Text Retrieval. In *First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 49–58, 1999.

[5] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In *TREC-2*, pages 57–66, 1994.

[6] N. Craswell, P. Bailey, and D. Hawking. Server Selection on the World Wide Web. In *Proc. AC Digital Libraries Conf.*, pages 37–46, 2000.

[7] T. E. Doszkocs. CITE NLM: Natural Language Searching in an Online Catalog. *Information Technology and Libraries*, 2:364–380, December 1983.

[8] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Comp. Linguistics*, 19(1):61–74, 1993.

[9] J. C. French and A. L. Powell. Metrics for Evaluating Database Selection Techniques. *World Wide Web*, 3(3), 2000.

[10] J. C. French, A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey, and Y. Mou. Comparing the Performance of Database Selection Algorithms. In *Proc. AC SIGIR '99*, pages 238–245, 1999.

[11] J. C. French, A. L. Powell, C. L. Viles, T. Emmitt, and K. J. Prey. Evaluating Database Selection Techniques: A Testbed and Experiment. In *Proc. SIGIR'98*, pages 121–129, 1998.

[12] N. Fuhr. A Decision-Theoretic Approach to Database Selection in Networked IR. *AC Trans. on Information Systems*, 17(3):229–249, 1999.

[13] S. Gauch, J. Wang, and S. M. Rachakonda. A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases. *AC Trans. on Information Systems*, 17(3):250–269, 1999.

[14] F. Gey, M. Buckland, A. Chen, and R. Larson. Entry Vocabulary – A Technology to Enhance Digital Object Search. In *Proc. of the First Inter. Conf. on Human Language Technology*, 2001.

[15] F. Gey, H. Jiang, A. Chen, and R. Larson. Manual Queries and Machine Translation in Cross language Retrieval and Interactive Retrieval at TREC-7. In *Text REtrieval Conf. (TREC-7)*, pages 527–539, 1999.

[16] F. C. Gey and A. Chen. Phrase Discovery for English and Cross-Language Retrieval at TREC-6. In *Text REtrieval Conf. (TREC-6)*, pages 637–648, 1998.

[17] F. C. Gey, A. Chen, J. He, L. Xu, and J. Meggs. Term Importance, Boolean Conjunct Training, Negative Terms, and Foreign Language Retrieval: Probabilistic Algorithms at TREC-5. In *Text Retrieval Conf. (TREC-5)*, 1996.

[18] L. Gravano and H. García-Molina. Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. In *Proc. of the 21st VLDB Conf.*, pages 78–89, 1995.

[19] L. Gravano, H. García-Molina, and A. Tomasic. GlOSS: Text-Source Discovery over the Internet. *AC Trans. on Database Systems*, 24(2):229–264, 1999.

[20] D. Harman. Towards Interactive Query Expansion. In *Proc. AC SIGIR'88*, pages 321–331, 1988.

[21] D. Hawking and P. Thistlewaite. Methods for Information Server Selection. *AC Trans. on Info. Systems*, 17(1):40–76, 1999.

[22] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proc. AC SIGIR '94*, pages 192–201, 1994.

[23] W. Hersh, S. Price, and L. Donohoe. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. In *Proc. of the 2000 American edical Informatics Association (A IA) Symposium*, 2000.

[24] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka. Overview of IR Tasks at the First NTCIR Workshop. In *The First NTCIR Workshop on Japanese Text Retrieval and Term Recognition*, pages 11–22, 1999.

[25] M. Kluck and F. Gey. The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval. In *Cross-language Information Retrieval Evaluation, Proc. of the CLEF 2000 Workshop*, 2001.

[26] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proc. AC SIGIR'97*, pages 324–332, 1997.

[27] W. Meng, K.-L. Liu, C. Yu, X. Wang, Y. Chang, and N. Rishe. Determining Text Databases to Search in the Internet. In *Proc. of the 24th VLDB Conf.*, pages 14–25, 1998.

[28] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. The Impact of Database Selection on Distributed Searching. In *Proc. AC SIGIR '00*, pages 232–239, 2000.

[29] B. Schatz, H. Chen, et al. Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurence Lists for Information Retrieval. In *Proc. AC Digital Libraries Conf.*, 1996.

[30] J. Xu and J. Callan. Effective Retrieval with Distributed Collections. In *Proc. AC SIGIR '98*, pages 112–120, 1998.

[31] C. Yu, W. Meng, K.-L. Liu, W. Wu, and N. Rishe. Efficient and Effective MetaSearch for a Large Number of Text Databases. In *Proc. AC CIK '99*, pages 217–224, 1999.

[32] B. Yuwono and D. L. Lee. Server Ranking for Distributed Text Retrieval Systems on Internet. In *Proc. of the Fifth Inter. Conf. on Database Systems for Advanced Applications*, pages 41–49, 1997.