The Potential to Improve Retrieval Effectiveness with Multiple Viewpoints

Allison L. Powell James C. French
Department of Computer Science *
University of Virginia
Charlottesville, Virginia 22903
{alp4g|french}@cs.virginia.edu

Technical Report CS-98-15

May 14, 1998

Abstract

We propose that providing multiple viewpoints of a document collection and allowing users to move among these viewpoints during a search or browse session will facilitate the location of useful documents. In this paper, we present the results of preliminary experiments that illustrate the potential benefits of this approach. These experiments utilize a full text view and a manually assigned topic descriptor view. We propose a measure for determining the amount of agreement among topic descriptors assigned to documents and show results for the TREC [7] SJMN and ZIFF collections. For those same collections, we then show the potential retrieval improvement when the two viewpoints are utilized.

1 Introduction

The information contained in collections of technical documents is often fully exploited by only the most experienced searchers. These searchers can draw upon years of general knowledge and knowledge about specific document collections to form effective queries. However, the task of finding useful documents can be frustrating for inexperienced searchers, novices to a field or researchers exploring document collections from a related field. These individuals can be unfamiliar with the contents and terminology of a document collection. This increases the difficulty of forming a suitable initial query and/or finding a good starting point for browsing the collection. Inexperienced searchers are also less likely to know effective strategies for reformulating queries or incorporating adjunct information. These individuals may be more likely to become "lost" while browsing the collection or may never locate interesting documents that are not directly connected in some way to their starting point.

We propose that providing multiple viewpoints of a document collection and allowing users to move among these viewpoints during a session will provide more effective access to collections. We propose that utilizing multiple views of a collection will enable both inexperienced and experienced searchers to more fully exploit the information contained in a document collection. Utilizing multiple views of a document collection will provide multiple opportunities to gain understanding of the material covered by the collection. In addition, users can refine searches and browse from the viewpoint that they find easiest to understand or move from viewpoint to viewpoint as they gain

*This work supported in part by DARPA contract N66001-97-C-8542 and NASA Graduate Student Researchers Program fellowship NGT5-50062.

understanding. This approach can support multiple searching styles. Finally, connections among documents may be more apparent in some views than in others; documents apparently unrelated in one view may show similarity in others. We propose that this aspect of using multiple viewpoints of a collection can be very effective when exploring a collection. Similarities revealed by an alternate view of the documents may serve to link groups of documents that use different terminology for the same concepts or may lead to interesting documents from a related discipline.

In this paper, we present the results of preliminary experiments that illustrate the potential of this course of research. We show that adjunct information about documents in our collection is assigned consistently and that this information can be exploited more effectively by considering it separately from the document text.

2 Related Work

Subject information has been used in other approaches, but not in the way that we are proposing. Srinivasan [12, 13] has conducted a series of experiments in which documents are represented by both full text and subject information. Given an initial query, subject terms from the highest ranking retrieved documents are added to the initial query. The reformulated query is then issued. However, no relevance judgements or user interactions are required so there is no guarantee that the topic information from the top-ranked retrieved documents captures the user's intentions. However, the results of Srinivasan's experiments illustrate the utility of utilizing topic information.

In the CHESHIRE system, Larson [9] employed the *classification clustering* approach to expand the vocabulary that could be used to access a bibliographic record. This approach clustered MARC bibliographic records by call number. Each call number cluster was indexed using all terms from the titles and Library of Congress Subject Headings (LCSH) of all records in the cluster. Therefore, a term or subject heading occurring in any document in a cluster could be used to retrieve that cluster. Cluster summaries, including the most frequently occurring subject headings were then displayed to a user; call numbers from user-selected clusters were incorporated into a reformulated query that returned individual records.

In addition to his description of the CHESHIRE system, Larson [9, p. 136-139] also provides an overview of studies that illustrate the problem of creating effective subject searches.

The library and information science literature contains a number of studies specifically investigating subject searching within online public access catalogs (OPACs). In general, subject searching in this context has been considered as a specific activity, not in conjunction with other types of searching. Drabenstott and Weller [5] note that lists of subject categories for browsing are often longer than a user is willing to deal with. However, users can have difficulty locating a topic using an exact subject search; only approximately 57 percent of the subject queries in their study exactly matched the controlled vocabulary of the library catalog). They have proposed an interface for guiding searchers to a correct subject descriptor, but do not support multiple subject descriptors for a single item.

The studies described above imply that while subject information can be used by a searcher to make decisions about potential document relevance, subject searching alone is generally not highly effective for finding documents about a particular subject area. We propose that our use of multiple views of a collection can help direct searchers to useful topic areas.

Research in query combination and data fusion has shown that using multiple document representations and/or multiple formulations of an information need can improve retrieval performance. Rajashekar and Croft [11] report a number of experiments using multiple query and documents representations in the INQUERY system. They performed experiments using single representations of queries with one, two or three representations of documents and two or three representations of queries with one document representation. They found consistent improvement when combined rep-

resentations were used. Belkin *et al.*[2] studied the effects of using multiple query representations. Independent experiments were performed by the authors for the TREC-2 conference. Both sets of experiments showed that combinations of query representations were beneficial. The results of the independent experiments were combined in a larger data fusion experiment. In general, the combinations outperformed the inputs.

The *classification clustering* functionality of CHESHIRE is still available in CHESHIRE II [10], but this is no longer the main focus. Instead, CHESHIRE II allows users to specify either Boolean queries with unranked results or natural language queries with probabilistically ranked results. Users can also specify both and see the merged results of both types of searches.

Ingwersen [8] argues for multiple representations of a user's cognitive state in addition to multiple representations of items in an information space.

We will show that the combination of results from two views of a collection produces results that are an improvement upon the results of either of the constituent views. This confirms that the behavior reported in [2, 11] can be observed when multiple views of the collection are used.

3 Motivation

There can potentially be many views of a document collection. For these experiments, we investigated a scenario in which there are two views of a collection based on terms within the documents and topic descriptors that have been assigned to the documents. This scenario was chosen because topic or subject information is a fairly common form of auxiliary information. Topic information can also be useful to searchers, prompting browsing if items are organized by topic (such as books on library shelves) or allowing searchers to eliminate obviously off-topic items.

An operational system that incorporates multiple views of a document collection will rely upon human interaction to initiate and focus transitions between views. The operational system will need to take into account both theoretically useful transitions and expected human strengths and weaknesses. For example, it has been shown that users can experience difficulty matching the restricted vocabulary of topic descriptors when constructing subject queries. Therefore, in our current proposed operational model, users will start with an initial query to the term-based view of the collection. Users will interact with that view of the collection to select a set of useful documents. The representations of these documents in the topic-based view will be used to form a query to the topic-based view of the collection. The user will be free to iterate between the views or within a single view.

While it is important to consider users during the construction of an operational system, conducting user studies to determine the basic feasibility of such an approach or to determine all of the baseline parameters of a prototype would be time consuming. Therefore, we have begun a series of preliminary experiments to gauge the feasibility of this approach, to determine the amount of exploitable information in the data and to identify effective transitions between views. In this paper, we present the results of some of these experiments.

4 Test Collections and Preparation

4.1 Test Collections

Our test collections were drawn from the TREC TIPSTER corpus. The goal of the Text REtrieval Conferences (TREC) is to "encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results." [1] TREC participants are furnished with (1) collections of documents, (2) sets of queries[1] that are statements of information needs (many of the queries are several paragraphs long), and (3) a relevance judgements file. For each query, the relevance judgements file contains a list of documents that a human judge

[1] In the TREC parlance, queries are referred to as "topics". To avoid confusion in this paper, we use the term queries.

has determined are relevant to the query (i.e. satisfies the information need expressed by the query). This enables participants to measure how well their experimental systems located documents that satisfied the queries.

4.2 Collection Preparation

All experiments were run on preprocessed versions of the TREC San Jose Mercury News (SJMN) and Ziff Publishing (ZIFF) collections. The SJMN collection consists of news articles that appeared in the San Jose Mercury News in 1991. The ZIFF collection consists mainly of computer-related articles published in Ziff Publishing publications from 1988 to 1992. However, the ZIFF articles located on TREC disk 3 contain a large number of short product-description articles [6]. While some of the product-description articles had been assigned topic descriptors, the articles themselves were of low quality for our experiments. There was generally very little text in the articles. The comparatively small size of these articles could make experimental results difficult to interpret, therefore the ZIFF articles on disk 3 were excluded from our test collection.

In addition to the full text of the articles, many of the documents in both collections contain manually assigned topic descriptors. Different sets of topic descriptors were used for the SJMN and ZIFF collections. These two sets represent two styles of topic descriptors. The topic descriptors used for the SJMN collection are generally one-word descriptors, (e.g. US; PRESIDENT; INCREASE; COST; PROGRAM). They are sometimes used in conjunction with one another to convey more complex concepts, so a topic descriptor can be repeated in a document (e.g. US; MILITARY; US; CONGRESS). The topic descriptors used for the ZIFF collection are phrase-based (e.g. Telecommunications; Long-distance telephone services; Reliability). There are instances of apparently accidental topic descriptor duplicates in the ZIFF collection.

To provide test collections in which the effects of the different views could be observed more easily, only documents with both article text and manually assigned topic descriptors were included in our preprocessed collections. As a result, all documents appear in both views. To simplify analysis, duplicate descriptors and empty descriptors were elided (e.g. US; CONGRESS; US would become US; CONGRESS while US; ; PRESIDENT would become US; PRESIDENT). There are 1635 unique topic descriptors assigned to documents in the SJMN collection and 8619 assigned to documents in the ZIFF collection.

Finally, initial experiments that included queries with only one relevant document artificially increased retrieval effectiveness in the topic descriptor view. As a result, only queries for which there are two or more relevant documents were included in the query evaluation set.

To differentiate our preprocessed collections and reduced query sets from the original TREC versions, we will refer to our collections as $SJMN'$ and $ZIFF'$ and the reduced query sets as $Q_{SJMN'}$ and $Q_{ZIFF'}$ respectively. See Table 1 for collection statistics.

	$SJMN$	$SJMN'$	$ZIFF$	$ZIFF'$
Num. docs	90,257	70,981	293,121	126,321
Num. queries	150	132	250	116

Table 1: Original and preprocessed collection statistics.

4.3 Relevant Document Topics (RDT) Sets

An important preliminary step for the experiments described below is the creation of Relevant Document Topics (RDT) sets. For each collection $COLL'$ being considered, an RDT set is constructed for each query q in $Q_{COLL'}$. The RDT set consists of $(document\ ID, topic\ descriptors)$ pairs for each of the documents in $COLL'$ judged relevant to q. So, for a collection $COLL'$,

$$\forall q \in Q_{COLL'},$$
$$RDT(q, COLL') = \{(ID(d), topics(d)) \mid d \in COLL' \wedge Rel(q, d) = 1\}$$

where $topics(d)$ are the topic descriptors assigned to document d and $Rel(q, d) = 1$ if document d has

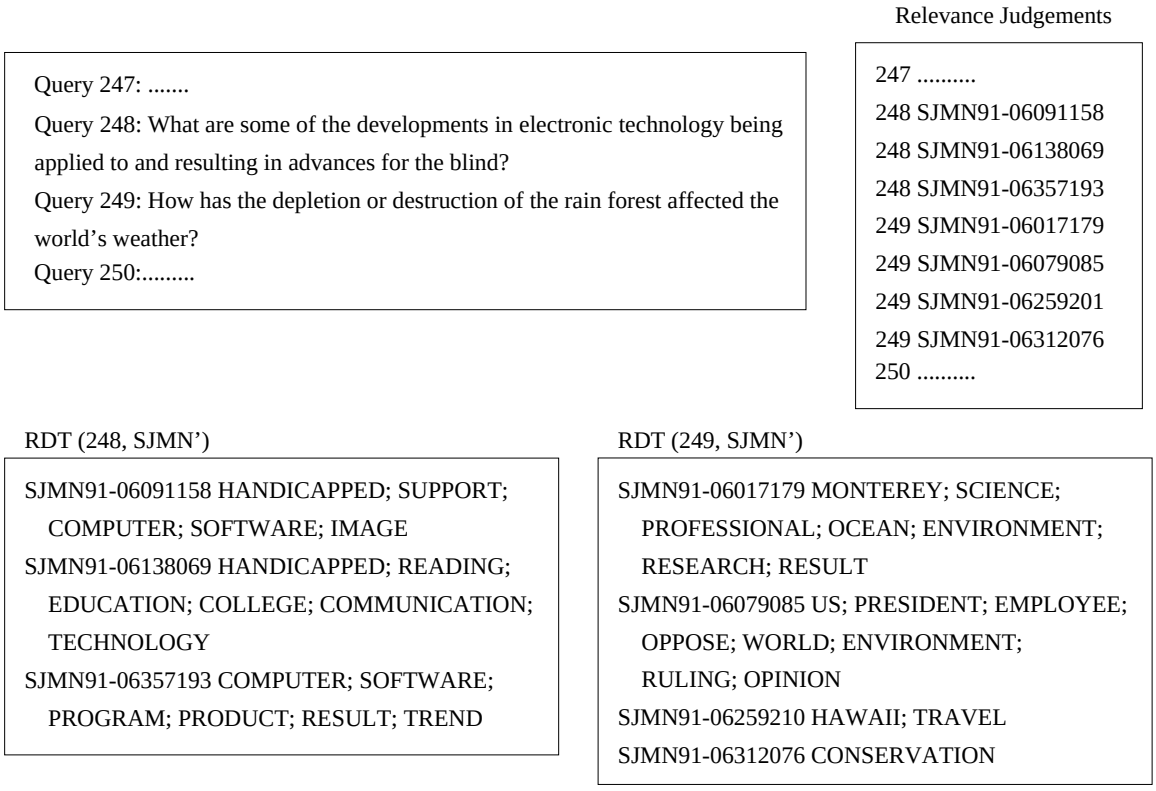

Figure 1: Creation of Relevant Document Topics (RDT) Sets

been judged relevant to query q.

Note that if a query occurs in the query set for more than one collection it will have more than one RDT value. Also, because elements of the sets are $(document\ ID, topic\ descriptors)$ pairs, a topic descriptor can appear multiple times in an RDT set.

Figure 1 is a simplified illustration of the creation of $RDT(248, SJMN')$ and $RDT(249, SJMN')$. Given the query identifiers 248 and 249, the set of document IDs of documents relevant to each query is extracted from the relevance judgements file. The document IDs and the topic descriptors assigned to the documents are then placed in the RDT set. The text of queries 248 and 249 is shown to provide context.

5 Test Collection Analysis

5.1 Consistency of Topic Descriptor Assignment

An initial question that must be answered before judging the results of an approach such as the multiple view approach for a collection is whether the collection is suitable for the approach. For example, if topic descriptors are assigned randomly to documents in a collection, the inclusion of those topic descriptors will degrade the retrieval performance. We have developed a measure to determine the degree consistency in topics assigned to a set of documents that are relevant to the same query (as opposed to topics assigned to the document collection at large).

It is not the case that all documents in an RDT set have the same topics assigned to them. In fact, it is rare for even a single topic descriptor to be assigned to all documents in an RDT set. In only

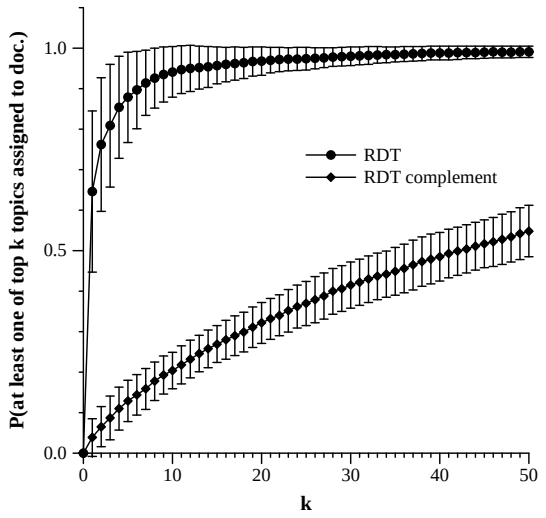

Figure 2: Topic Consistency CDF ($SJMN'$)

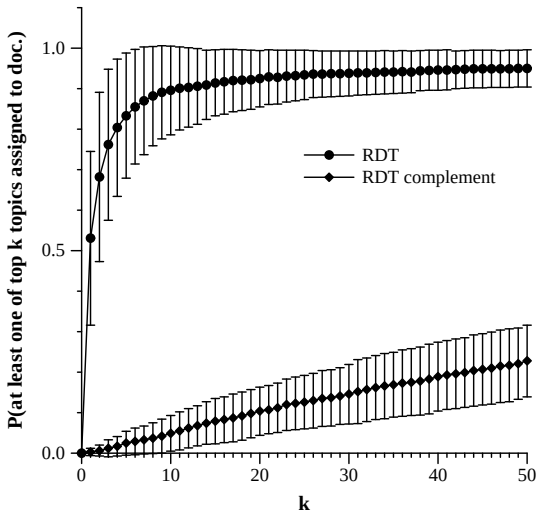

Figure 3: Topic Consistency CDF ($ZIFF'$)

7 of the 132 RDT sets for $SJMN'$ and 10 of the 116 RDT sets of $ZIFF'$ did any one topic descriptor occur in all documents in the set. Except for one RDT set with 5 items and another with 12, all sets with a topic occurring in all items had 2 or 3 items.

This difference could be caused by lack of consistency among different indexers responsible for assigning the topic descriptors. It could also be the case that even though all of the documents in an RDT set are relevant to the same query, the documents have different nuances that lead to different topic descriptors. Therefore, it is illuminating to consider a looser definition of consistency. Instead of expecting an exact match of topics assigned to documents in an RDT set, we measure the percentage of documents in an RDT set that contain at least one of the top k topics.

For each document collection and each RDT set for that collection, we first extract all topic descriptors assigned to a document in the RDT set. The topic descriptors are then sorted in decreasing order of occurrence (i.e. topic descriptors assigned to the most documents in the RDT set are considered first). Ties are broken essentially randomly, based partially on the order in which a topic descriptor was encountered. Starting with the most frequently occurring topic descriptor, then considering the next most frequently occurring, etc., we measure the percentage of documents in the RDT set to which that descriptor or any of the ones before it were assigned.

We then considered the complement of each RDT set for each document collection (i.e. all documents in the collection that are not relevant to the query). We calculated the same measures using the topic descriptors extracted from the corresponding RDT set to characterize the way in which the topic descriptors from the RDT set were assigned to the collection as a whole.

For example, consider $RDT(248, SJMN')$ from Figure 1. In this example, the RDT set would contain documents SJMN91-06091158, SJMN91-06138069 and SJMN91-06357193. The complement of the RDT set would contain all documents from $SJMN'$ *except* SJMN91-06091158, SJMN91-06138069 and SJMN91-06357193. The topic descriptors would be considered in the order shown in Table 2 since HANDICAPPED, COMPUTER and SOFTWARE occur in two documents from the set and the remaining topic descriptors occur in one document each. We would calculate the fraction of documents in the RDT set to which HANDICAPPED was assigned, then COMPUTER or HANDICAPPED, etc. (see the third column of

Table 2). Then we would calculate the fraction of documents in the complement of the RDT set to which HANDICAPPED, COMPUTER, etc. were assigned.

Term	Num. Docs	Cumulative fraction
HANDICAPPED	2	0.667
COMPUTER	2	1.0
SOFTWARE	2	1.0
SUPPORT	1	1.0
IMAGE	1	1.0
READING	1	1.0
EDUCATION	1	1.0
COLLEGE	1	1.0
COMMUNICATION	1	1.0
TECHNOLOGY	1	1.0
PROGRAM	1	1.0
PRODUCT	1	1.0
RESULT	1	1.0
TREND	1	1.0

Table 2: Example calculation using RDT set.

The results for collection $SJMN'$ and $ZIFF'$ are displayed as averaged cumulative distribution functions in Figures 2 and 3. For each collection, and for each value of k, the percentage of items to which at least one of the top k topics has been assigned, averaged over all RDT sets (RDT complements) is plotted. Error bars are plotted at $\pm$ one standard deviation.

Note that for both collections, the most frequently occurring topic is assigned to 50-65% of the items in the RDT sets and less than 5% of the items in the RDT complement sets on average. There is rapid improvement in the RDT curve up to $k = 6$ or $k = 7$, while the RDT complement curve shows a steady gain. This suggests that for small values of k, topic descriptors could be utilized for retrieval without bringing in a very large number of extraneous documents. Beyond approximately $k = 10$, we see diminishing returns in the RDT curve and the same steady increase in the RDT complement curve.

This implies that there is consistency in the assignment of topic descriptors to documents that are relevant to the same query compared to documents in the collection at large. This also suggests that it would be beneficial to incorporate topic descriptors when attempting to locate relevant (or additional relevant) documents. It is encouraging that a large benefit can be gained from the use of a relatively small number of topic descriptors, decreasing the potential burden on a user for selecting these descriptors.

5.2 Potential Retrieval Improvement

Our next experiment was to verify that the combination of results from the full text and topic views is an improvement upon the results of either of the constituents individually.

For this experiment, the two views of the document collections are represented by two different indexes of each collection. These indexes were constructed with SMART version 11[2] using document indexing parameters `lnc` and query indexing parameters `ltc`. These parameters determine the weights of features in the indexes; these specific parameters have been shown to work well for TREC collections [3, 4]. Additional information about the parameters is available in the SMART documentation. The full text index was constructed by indexing all available document information for the documents in $SJMN'$ and $ZIFF'$ (including the topic descriptor field). The topic descriptor index was constructed by indexing only the topic descriptor field of each document. The multi-word $ZIFF'$ topic descriptors were treated as phrases, not decomposed into individual terms.

From the preprocessing steps of the previous experiment, we had available the sets of queries $Q_{SJMN'}$ and $Q_{ZIFF'}$. We constructed a corresponding topic query for each term query in $Q_{SJMN'}$ and $Q_{ZIFF'}$. We refer to the topic query sets as $TQ_{SJMN'}$ and $TQ_{ZIFF'}$, respectively. These topic queries assumed "perfect" information of the topics assigned to documents relevant to the corresponding term query. Each topic query was constructed by concatenating the topic descriptor portions of the information contained in the corresponding RDT set. For example, referring again to Figure 1, topic query 248 for collection $SJMN'$ would be

[2]SMART version 11.0 is available via anonymous ftp from ftp://ftp.cs.cornell.edu/pub/smart/

```
HANDICAPPED; SUPPORT; COMPUTER; SOFTWARE;
   IMAGE
HANDICAPPED; READING; EDUCATION;COLLEGE;
   COMMUNICATION; TECHNOLOGY
COMPUTER; SOFTWARE; PROGRAM; PRODUCT;
   RESULT; TREND
```

Note that $ZIFF'$ would have a different topic query 248.

After creating the full text and topic view indexes of $SJMN'$ and $ZIFF'$, we ran four retrieval runs:

1. collection $SJMN'$ full text view, query set $Q_{SJMN'}$;
2. collection $ZIFF'$ full text view, query set $Q_{ZIFF'}$;
3. collection $SJMN'$ topic view, query set $TQ_{SJMN'}$;
4. collection $ZIFF'$ topic view, query set $TQ_{ZIFF'}$.

For each retrieval run, we had SMART return the top 100 documents for each query in terms of similarity. These results were evaluated using the SMART `tr_eval` function. We then extracted the retrieved documents and similarities for each of the four result sets.

The information extracted from the result sets was used to create the merged results. We performed a very simple merge of the full text and topic descriptor results for both $SJMN'$ and $ZIFF'$. For each collection, if a document was retrieved by both query q_i *and* corresponding topic query tq_i, the two similarity values were added to form the similarity value in the merged result set. If a document was retrieved by only query q_i *or* corresponding topic query tq_i then the single similarity value was simply carried forward to the merged result set. We then used the `trec_eval` script that was distributed with the TREC data to evaluate the performance of the merged result set (which could contain up to 200 documents).

The three recall-precision curves are plotted for $SJMN'$ and $ZIFF'$ in Figures 4 and 5. Recall is the percentage of documents relevant to a query that are retrieved in response to that query. Precision is the percentage of documents retrieved in response to a query that are relevant. For a given query, precision values are calculated at fixed recall points. The recall-precision curves in Figures 4 and 5 represent an average of the curves for sets of queries. Note that in both figures, the merged results represent a large improvement over the full text results (which also indexed the topic descriptor field). This suggests that the topic information can be more fully exploited by considering it separately.

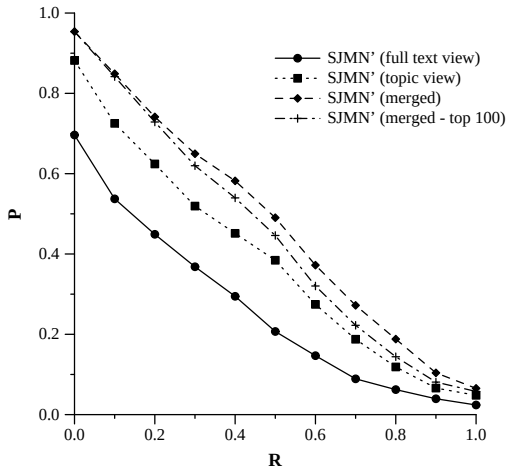

Figure 4: Retrieval improvement ($SJMN'$)

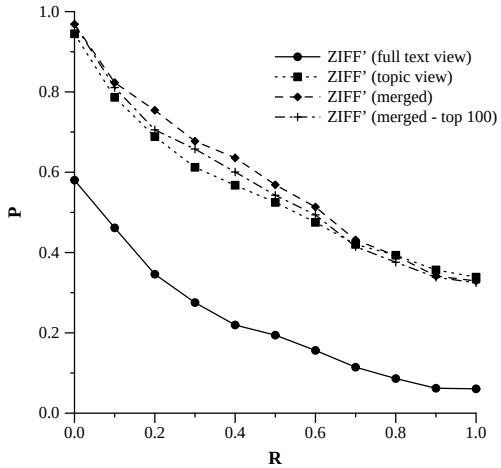

Figure 5: Retrieval improvement ($ZIFF'$)

One could argue that the merged results set has an unfair advantage in this comparison because that set could contain up to 200 retrieved documents for each query. The ability to retrieve documents that only show similarity to the query in one viewpoint is a strength of the multiple view approach; this could be hampered by restricting the number of documents considered. To determine the impact of the larger number of documents, we evaluated the retrieval results using the top 100 documents for each query in the merged set. These results are shown on Figures 4 and 5 as the "merged - top 100" plot. These results show a small decrease in performance when compared to the original merged results; however the performance is still good.

We realize that this is a rather artificial test; in an operational environment this "perfect" topic information will not be available. In fact, in a working document collection without prepared queries and relevance judgements files, this "perfect" topic information is unknowable. However, we do feel that the merged results represent a reasonable heuristic for potential performance improvements.

6 Conclusions and Future Research

These experiments have shown that in a document collection for which there is correlation between document relevance and adjunct topic information, it is more effective to treat this topic information separately than to include it with the full document text. However, these preliminary experiments have only illustrated the potential of the approach; we have made assumptions that cannot be applied directly to an operational system. The challenge now is to determine to what degree this potential can be realized in a more realistic environment.

Specifically, we must determine if we can achieve improvements similar to these using only partial topic information. If so, how much topic information will be required and how should that information be provided by the user?

In addition, in an operational system, we must determine what results are displayed to users. As users interact with a system such as the one partially described above, they can move from one view to another multiple times. They can therefore potentially generate multiple sets of documents to be examined. An issue is how to display these sets of documents in a way that helps to enhance the users' understanding of the information covered by the document collection while also enhancing their understanding of the impact of the selections made in previous iterations.

References

[1] Text REtrieval Conference (TREC) Overview. http://trec.nist.gov/overview.html, version last updated February 5, 1998.

[2] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management*, 31(3):431–448, 1995.

[3] Chris Buckley, James Allan, and Gerard Salton. Automatic routing and ad-hoc retrieval using SMART : TREC 2. In D. K. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*, pages 45–56. NIST Special Publication 500-215, March 1994.

[4] Chris Buckley, Gerard Salton, and James Allan. Automatic retrieval with locality information using SMART. In D. K. Harman, editor, *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 59–72. NIST Special Publication 500-207, March 1993.

[5] Karen M. Drabenstott and Marjorie S. Weller. The Exact-Display Approach for Online Catalog Subject Searching. *Information Processing & Management*, 32(6):719–745, November 1996.

[6] 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. of the 21st ACM SIGIR Conference on Information Retrieval*, Melbourne, Australia, 24-28 August 1998. (to appear).

[7] Donna Harman. Overview of the Fourth Text Retrieval Conference (TREC-4). In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, Gaithersburg, MD, 1996.

[8] Peter Ingwersen. Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *Journal of Documentation*, 52(1):3–50, March 1996.

[9] Ray R. Larson. Classification clustering, probabilistic information retrieval and the online catalog. *The Library Quarterly*, 61(2):133–173, April 1991.

[10] Ray R. Larson, Jerome McDonough, Paul O'Leary, Lucy Kuntz, and Ralph Moon. Cheshire II: Designing a Next-Generation Online Catalog. *Journal of the American Society for Information Science*, 47(7):555–567, July 1996.

[11] T. B. Rajashekar and W. Bruce Croft. Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society for Information Science*, 46(4):272–283, May 1995.

[12] Padmini Srinivasan. Optimal Document-Indexing Vocabulary for MEDLINE. *Information Processing & Management*, 32(5):503–514, 1996.

[13] Padmini Srinivasan. Query Expansion and MEDLINE. *Information Processing & Management*, 32(4):431–443, 1996.