Document Retrieval Using
The MPS Information Server
(A Report on the TREC-4 Experiment)

François Schiettecatte
Valerie Florance, Ph.D.
(francois@fsconsult.com, vf@fsconsult.com)
FS Consulting
1890 Highland Avenue, Rochester, NY, 14618
http://www.fsconsult.com

1 Introduction

This paper summarizes the results of the experiments conducted by FS Consulting as part of the Fourth Text Retrieval Experiment Conference (TREC-4). FS Consulting participated in Category C, ran the ad hoc experiments, and produced two sets of results, fsclt1 and fsclt2. Our long-term research interest is in building systems that help users find information to solve real-world problems. Our TREC-4 participation centered on three goals: to model information seeking behavior of an average searchers; to refine a retrieval system for use in a real-life setting; and to build tools to help searchers retrieve information effectively from large databases. Our two TREC-4 experiments were designed around a model of an experienced end user of information systems, one who might regularly use a system like the MPS Information Server while seeking information in a workplace or library setting. Our TREC-4 experiments provided us with baseline experience for initiating future retrieval projects and for participation in the TREC-5 interactive track.

2 Overview of FS Consulting TREC-4 Experiments

In the TREC-4 experiments we set out to answer two questions:

- How does our system perform on large databases using queries constructed by an experienced end-user/searcher?

We began with the assumption that our information seeker had previous experience using online retrieval systems commonly available in libraries (e.g., online library catalogs and bibliographic systems like MEDLINE). Although the search interface varies considerably, most library systems default to a novice-type search interface that allows a searcher to enter one or two terms and apply a Boolean operator to relate them. To aid the more advanced end-user/searcher, academic libraries typically provide search aids and training sessions that teach searchers to construct more complicated Boolean statements and employ controlled vocabularies for term selection. Additionally, information seekers who patronize librarian-mediated search services have observational experience to draw upon: often, they are asked to create query statements, or to work directly with librarians performing the search as they construct and revise search statements. It is this combination of context and experience that constitutes the background knowledge of our experienced searcher.

For our initial experiment, fsclt1, we allowed only manually-constructed queries, constructed offline. The searcher was permitted to employ any number of terms into simple or complex Boolean arguments. For this experiment, a single user query was entered for each topic, and a relevance ranked output was generated for each, using standard system features of the MPS Information Server.

- Can relevance ranking be improved within a retrieved set?

Studies and in-the-field observation of information seeking behavior demonstrate a number of important characteristics of real-world information seekers:

- they seek information to solve a problem;
- they want only enough relevant documents to answer their question;
- they don't want to review long lists of documents;
- their own relevance assessments are not absolute.

In our model of the experienced searcher, the information seeker is assumed to have an inexplicit mental model that shapes the size, shape and texture of a satisfactory search outcome. Although a system may retrieve 1000 documents in response to this searcher's query, s/he is not likely to review more than 10% of those before changing strategies, declaring the problem solved, or abandoning the search. Thus, we make the assumption that the best items must appear in the first 100 documents if they are to be viewed at all.

For our second experiment, fsclt2, we tested a tool that could help move the best documents to the top of an existing retrieved set. Reasoning that useful documents might be scattered throughout an initial retrieval set, we designed an automated query expansion feature to work within that set. For this experiment, the retrieval result set from each query in our first experiment was re-ordered using a relevance feedback algorithm.

3 Searcher Model and Guidelines

Because fsclt1 and fsclt2 employed the same query formulations, the same searcher model and guidelines apply to both. All query statements for the experiments were constructed by one person. The initial parameters of the searcher 'model' were defined as follows:

- s/he regularly searches on-line catalogs and bibliographic databases in an academic setting;
- s/he may have some search training, but is not a professional searcher;
- s/he dislikes reviewing large search outputs;
- s/he is seeking information to solve a real-life problem;
- s/he may not be a content expert in the topic area of a given question.

3.1 Instructions to Searcher

The following instructions guided query formulation:

- prepare a single search statement that will capture the most relevant documents for a given topic;
- use single or multiple terms, employing wild card capability to capture multiple versions of a word, and/or quotes around several words (e.g., "cardiac arrest") to create a fixed phrase;
- apply boolean logic as desired, using AND, OR or NOT operators. Create nested statements using parentheses if desired;
- consult the stop word list if needed, but no other databases are available for consultation;
- the total time taken to prepare a single query should not exceed 5 minutes.

3.2 Searcher Training

In preparation for the experiment, the searcher performed training exercises using the TREC-4 training data. First, general capabilities of the system and features of the search engine were described. Then, three topics (Christian coalition, prison overcrowding, oil spills) were selected from the training topics by the searcher. For each topic, a query formulation constructed by the searcher was run against the test database. Results were analyzed using the Trec Eval program. Lists of document headlines were provided to the searcher for examination. The searcher was allowed to reformulate and re-run training queries as many times as she desired. The search interface for training exercises employed a custom client application created for this experiment and MacWais, a freely available WAIS client for the Macintosh.

4 System Configuration

The MPS Information Server is a commercial full-text retrieval system that runs on a large number of Unix based platforms. Given a user query, the MPS system returns a list of relevance-ranked documents from a database. The system is capable of performing simple or complex term or phrase searching using parentheses, wildcards and Boolean operators. Soundex and fielded searching are also supported. The system is designed to favor precision over recall when performing searches. Because it supports a number of different protocols, including WAIS-88, Z39.50-V2, HTTP and Gopher+, the MPS Information Server is capable of responding to search requests from a wide variety of clients applications.

The TREC-4 experiments employed version 2.0 of the MPS Information Server running on a SparcStation 5/70. Four gigabytes of disk space were set aside, 2 GB for the TREC data and 2 GB for indices. For the purposes of the experiments, we built a custom client application. Running on an Apple Macintosh under A/UX 3.0, the custom client communicated with the MPS Information Server using the WAIS-88 protocol. This client application was designed to read TREC topic files, build a query by extracting a specific field (or fields) from the individual topic entries, run the queries against the server and save the query results in the TREC result format to a specified file. The results files could then be processed by the Trec Eval program to obtain the precision-recall values for the run.

A special parser was built to index the TREC database. We chose to extract the document ID field (the <DOCNO>) and the document title (where a title was available) to create a document headline for each document. The rest of the document was indexed as plain text, with the SGML tags extracted from the text, and the words stemmed using a plural stemmer. No additional information was extracted apart from the word positions to allow phrase and proximity searching. All keywords in the news articles were ignored, as required by the instructions.

While the MPS system's indexer starts up with an initial stop-word list (containing 377 words), it can choose to convert a word to a stop word if that word's total occurrence in the database reaches a specified value. The stop word value, which is site- and collection-dependent, would typically be set anywhere in the range of 20,000 to 50,000 occurrences. For the TREC-4 experiments, it was set at 150,000 occurrences to retain as many words as possible in the database. This resulted in a final stop-word list of 626 words.

Two databases were created, one containing disks 1 and 2 (the ad hoc training database) and the other containing disks 2 and 3 (the ad hoc test database). Each database took approximately 14 hours to build; their index sizes were about 800 MB each.

5 TREC-4 Results for FS Consulting Experiments

5.1 Experiment fsclt1

In the first experiment, the searcher created a single written query for each of 50 ad hoc topics. The queries were entered into the system and processed as a batch, without further involvement of the searcher. Search results were saved to a TREC-formatted result file.

5.1.1 Searcher Performance

Training exercises influenced the searcher's query formulation behavior in the following ways:

- she mistrusted the wild-card capability, preferring to enter multiple forms of a word;
- she was reluctant to add long lists of multiple synonyms, believing that they would dilute search results;
- she tried work-arounds to avoid the stop-list for common words like 'states' and 'united' (e.g., using "United States" as a bounded phrase).

The following examples are typical formulations used for fsclt1. The searcher wrote out the formulations, which were entered into the system by the researcher (FS) without further discussion or modification.

Topic 202: (nuclear AND treaty) AND ((violate OR violation) OR (Monitor OR monitoring))

Topic 212: ("intellectual property" OR copyright OR trademark OR patent) AND violation

Topic 217: (extraterrestrial OR extra-terrestrial) AND (life OR intelligence)

Topic 218: "steel mill" AND ("steel industry" OR (steel AND production))

Topic 221: ((drug OR gang) AND warfare) AND (prevent or prevention)

Topic 224: ("high blood pressure" OR hypertension) AND (treat OR treatment)

Topic 228: (pollution AND (reduce or reduction OR abate OR abatement)) OR (environment AND (recover OR recovery OR improve OR improvement))

Topic 231 ("National Endowment for the Arts" OR NEA) AND (appropriate OR appropriation OR budget OR fund OR funding)

Topic 241: ("professional conduct" OR malfeasance) AND (lawyer OR doctor OR physician OR attorney)

Topic 244: ("United States" OR U.S. OR America) AND Japan AND "trade balance"

Topic 248: (blind OR blindness) AND (electronic OR technology)

Most query formulations for fsclt1 employed parentheses and the AND and OR Boolean operators. As the examples indicate, not all capabilities of the system were employed (e.g., wildcard, soundex and "NOT" operator were not used for example). The bounded phrase was the most common special feature used. Missing parentheses in the examples above suggest that logic statements were not always constructed properly.

5.1.2 Server Performance for fsclt1

The results for fsclt1 produced the following precision/recall figures over all of the topics:

```
Queryid (Num):       all  fsclt1
Total number of documents over all queries
    Retrieved:    11472
    Relevant:      6501
    Rel_ret:       1103
Interpolated Recall - Precision Averages:
    at 0.00       0.5858
    at 0.10       0.3066
    at 0.20       0.2348
    at 0.30       0.2096
    at 0.40       0.1753
    at 0.50       0.1146
    at 0.60       0.0513
    at 0.70       0.0250
    at 0.80       0.0204
    at 0.90       0.0000
    at 1.00       0.0000
Average precision (non-interpolated) over all rel docs
                  0.1303
Precision:
  At    5 docs:   0.3918
  At   10 docs:   0.3571
  At   15 docs:   0.3306
  At   20 docs:   0.3122
  At   30 docs:   0.2810
```

```
  At  100 docs:   0.1524
  At  200 docs:   0.0913
  At  500 docs:   0.0425
  At 1000 docs:   0.0225
R-Precision (precision after R (= num_rel for a query) docs
retrieved):
    Exact:        0.1840
```

Overall, for all topics, 20% of the relevant documents were retrieved from the database and only 10% of the documents retrieved were relevant. Analysis of the distribution of relevant documents retrieved for each topic revealed that most were concentrated in the top 100 documents retrieved for each topic. Table 1 below shows R-precision scores for fsclt1's topics as compared to all TREC-4 participants.

Table 1: R-precision scores for fsclt1 at 100 and 1000 documents

R-precision Scores	@ 100	@ 1000
Above Median	15	4
At Median	4	8
Below Median	9	9
Poor	15	21
Worst	6	7
Average precision, 49 topics: .1303 • R-precision, 49 topics: .1840		

At the 100-document level, the MPS Information Server did reasonably well: 19 of the topics were at the median or above it. Topics 216,224, 226, 232, 234, 235, 242 and 245 showed 'above Median' R-precision scores at 100 documents. At 1000 documents, most of the fsclt1 results fall below the median. It should be noted, however, that only 5 out of the 49 searches retrieved 1000 documents.

Figures 1 and 2 below summarize differences in number of relevant documents retrieved for each topic in fsclt1, as compared to the overall TREC-4 median scores. Figure 1 demonstrates variation from the TREC-4 median at 100 documents; Figure 2 does the same at 1000 documents.

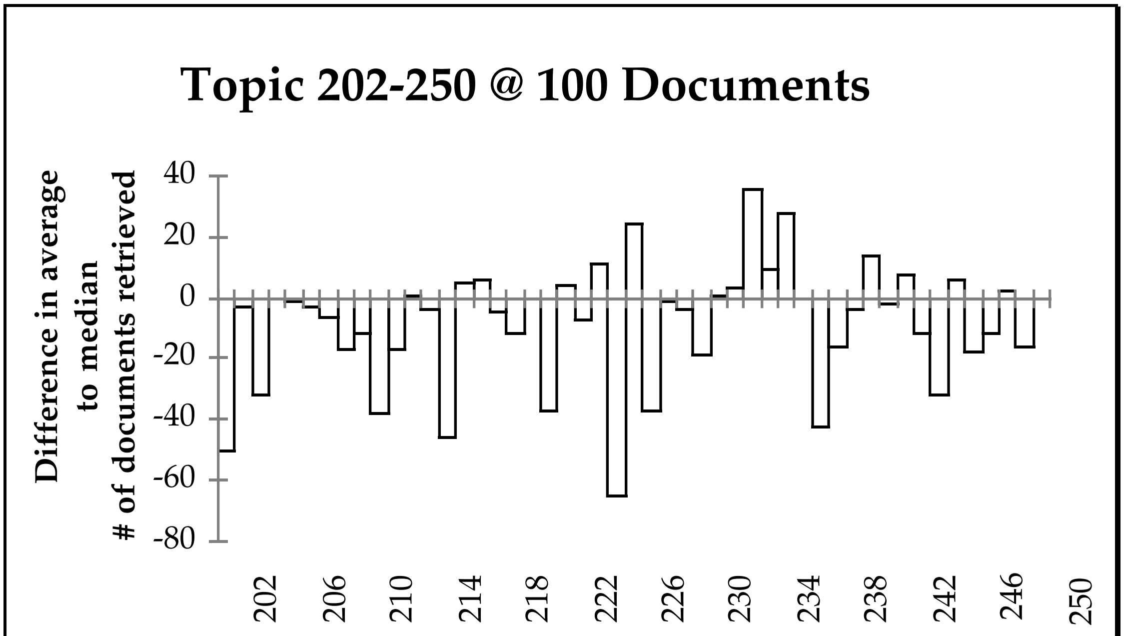


Figure 1

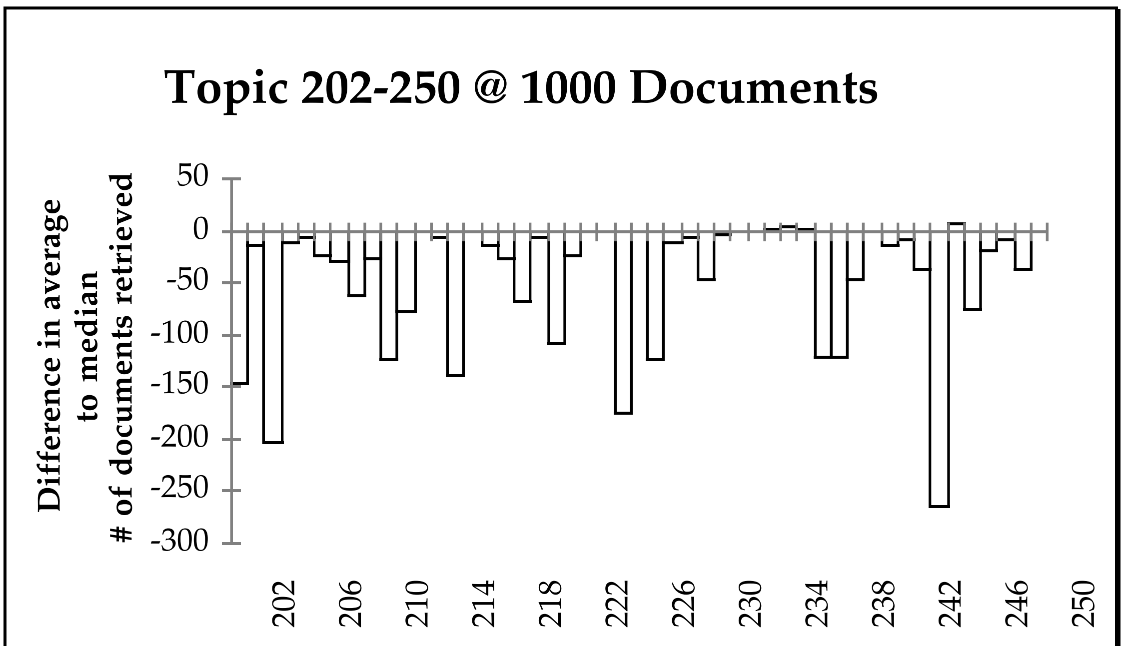


Figure 2

Figure 3 summarizes the overall variation from median R-precision scores for each topic in fsclt1.

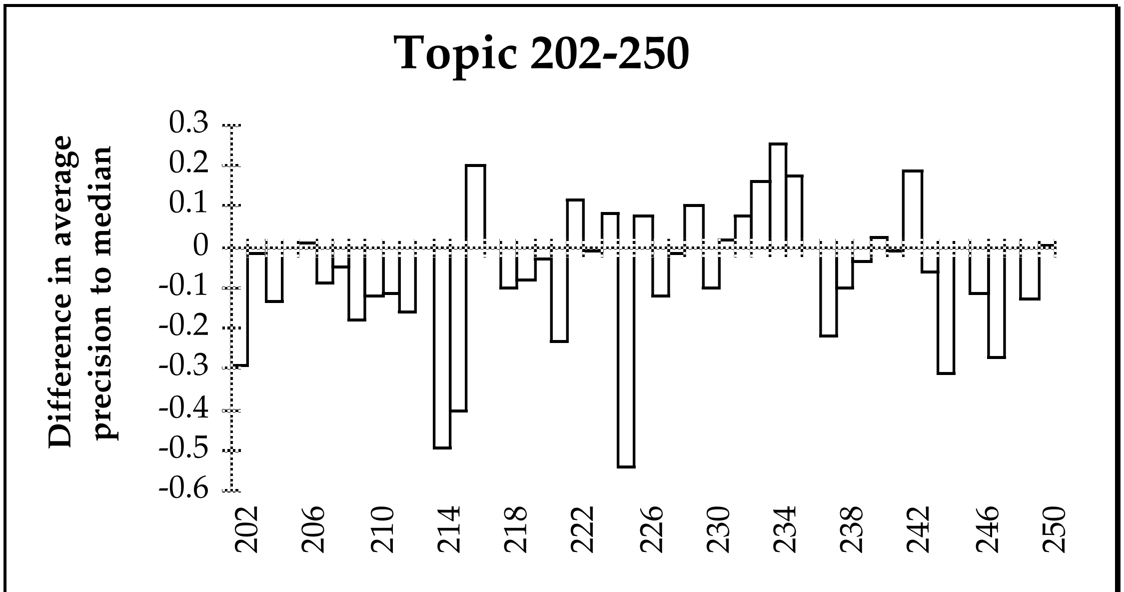


Figure 3

5.1.3 Success/Failure Analysis for fsclt1

Examination of results from fsclt1 centered on the most and least successful topic results. One important point that emerged immediately was the effect of the search formulations on retrieval success. Successful queries (i.e. those, whose results fell at or above the TREC-4 median scores) used fewer terms and relatively simple Boolean arguments. Results are consistent with the MPS Information Server's design that favors precision over recall. When fewer than 50 relevant documents were retrieved, half of them were ranked in the first 20 documents. When more than 500 documents were retrieved for a topic, all the relevant documents were ranked in the first 100 documents

5.1.3.1 Success analysis for fsclt1

Review centered on topics for which results were at median or better when compared to the overall outcomes for all TREC-4 participants. Topic 216, on osteoporosis, was one of the most successful outcomes in fsclt1. The query was formulated as *osteoporosis AND ((prevent OR prevention) OR (treat OR treatment)).*

The precision/recall figures for this topic were as follows:

```
Queryid (Num):       216  fsclt1
Total number of documents over all queries
    Retrieved:       82
    Relevant:        36
    Rel_ret:         29
Interpolated Recall - Precision Averages:
```

```
    at 0.00       1.0000
    at 0.10       1.0000
    at 0.20       1.0000
    at 0.30       1.0000
    at 0.40       0.8333
    at 0.50       0.7200
    at 0.60       0.6875
    at 0.70       0.5417
    at 0.80       0.3973
    at 0.90       0.0000
    at 1.00       0.0000
Average precision (non-interpolated) over all rel docs
                  0.6439
Precision:
  At    5 docs:   1.0000
  At   10 docs:   1.0000
  At   15 docs:   0.8667
  At   20 docs:   0.7500
  At   30 docs:   0.7000
  At  100 docs:   0.2900
  At  200 docs:   0.1450
  At  500 docs:   0.0580
  At 1000 docs:   0.0290
R-Precision (precision after R (= num_rel for a query) docs
retrieved):
    Exact:        0.6111
```

Note that while most of the relevant documents were retrieved (29 out of 36), only 82 documents were retrieved for the whole query. This produced a small results set for the user to look at, with most of the relevant documents concentrated at the top of the result set. All of the top 10 documents were relevant, as were 13 of the top 15, 15 of the top 20, and 21 of the top 30. This topic was most successful in terms of both the experiment's goals and the stated parameters of the searcher model.

5.1.3.2 Failure Analysis for fsclt1

Topics 208, 211, 214, 215, 243, 249 had a 'worst' R-precision score at 100 documents when compared to overall TREC-4 results. Examples of query formulations for 4 of these topics follow:

Topic 211: *((DWI OR drunk) AND (drive OR driving) AND (penalties)) AND mortality*
Topic 214: *"self-induced hypnosis"*
Topic 215: *"infant mortality rate" AND ("United States" OR U.S. OR America) AND (higher OR lower)*
Topic 249: *weather AND "rain forest" AND (deplete OR destroy OR disappear)*

To explore the effect of query construction on the retrieval outcome, post TREC-4 experiments were run (see section 6.1 below) in which several queries were reformulated by dropping one or more phrases. In every case, the retrieval scores improved.

5.2 Experiment fsclt2

The second experiment centered on re-ordering relevant documents within a retrieved set. We began by running each unedited query from fsclt1 against the TREC-4 database. For each result set, the first three documents in the set were flagged as being most relevant. Up to 10 terms were selected from each document and, of those, the most 'interesting' terms were chosen for automatic query expansion. Those terms were added to the original query with an exclusive OR and assigned weights equal to 10% of those assigned to original query terms. The query was run again and the results were saved to a TREC format result file. The searcher who formulated the original queries was not involved in this experiment.

The relevance feedback algorithm employed for query expansion in this experiment works by ranking all terms in selected documents by frequency of occurrence. The ten most 'interesting' terms were chosen from that list for further use. The most 'interesting' terms were defined as neither the most frequent or infrequent terms. Rather, frequency parameters were specified to eliminate the high and low ends of the spectrum. The experiment's final result sets were produced by expanding each original query to include new terms, assigning weights to the old and new terms, and re-ordering documents in the set based on new relevance weights. In no case were additional documents added to the original set.

This automated query expansion feature was designed as a tool that could be used by information seekers who, having retrieved a large set from an initial search, wish to increase the likelihood that all relevant documents retrieved were listed in the first 30 or 40 titles in the output list.

5.2.1 Server Performance for fsclt2

Experiment 2 produced the following precision/recall figures for all the topics:

```
Queryid (Num):        all   fsclt2
Total number of documents over all queries
    Retrieved:    11472
    Relevant:      6501
    Rel_ret:       1108
Interpolated Recall - Precision Averages:
```

```
    at 0.00       0.5830
    at 0.10       0.3028
    at 0.20       0.2275
    at 0.30       0.2012
    at 0.40       0.1673
    at 0.50       0.1077
    at 0.60       0.0465
    at 0.70       0.0238
    at 0.80       0.0206
    at 0.90       0.0000
    at 1.00       0.0000
Average precision (non-interpolated) over all rel docs
                  0.1248
Precision:
  At    5 docs:   0.3755
  At   10 docs:   0.3510
  At   15 docs:   0.3238
  At   20 docs:   0.3112
  At   30 docs:   0.2796
  At  100 docs:   0.1504
  At  200 docs:   0.0888
  At  500 docs:   0.0419
  At 1000 docs:   0.0226
R-Precision (precision after R (= num_rel for a query) docs
retrieved):
    Exact:        0.1826
```

These results are not very different from fsclt1 results. Table 2 shows R-precision scores for fsclt2 topics compared to all TREC-4 participants.

Table 2: R-precision scores for fsclt2 at 100 and 1000 documents

R-precision Scores	@ 100	@ 1000
Above Median	14	4
At Median	4	7
Below Median	9	10
Poor	14	18
Worst	8	10
Average precision, 49 topics: .1248 • R-precision, 49 topics: .1826		

For the fsclt2 experiment, more relevant documents should have appeared in the first 100 documents for query topics. As this table indicates, that did not happen, and results at 1000 documents showed unpredicted repatterning.

Figures 4 and 5 below summarize differences in the number of relevant documents retrieved with respect to the TREC-4 overall median at 100 documents and at 1000 documents respectively. With few exceptions, data are identical to those presented in Figures 1 and 2.

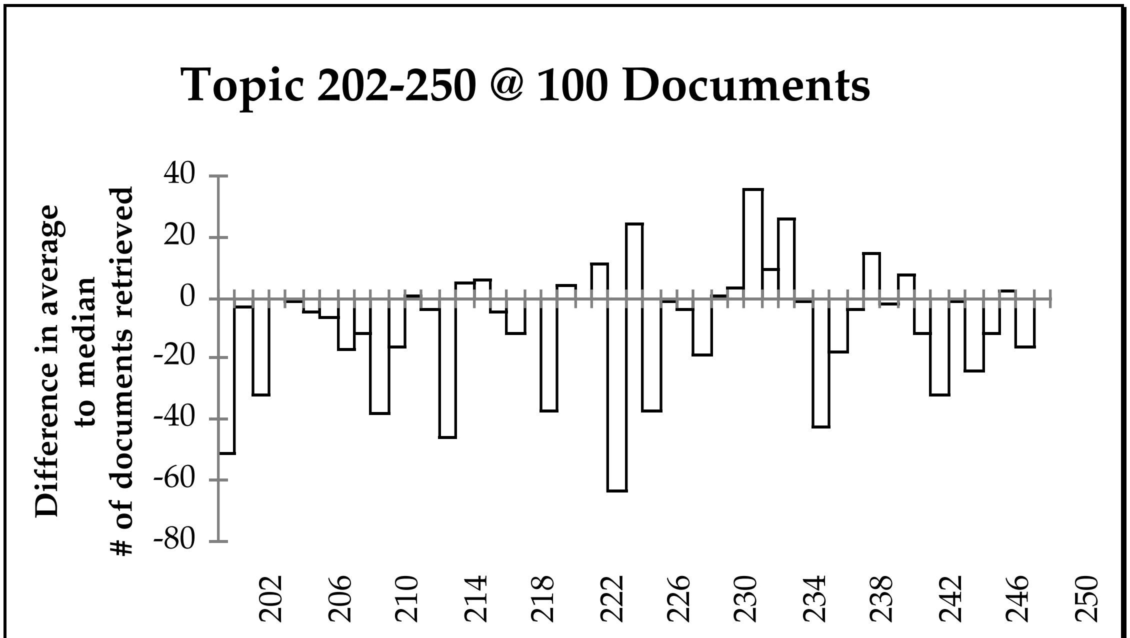


Figure 4

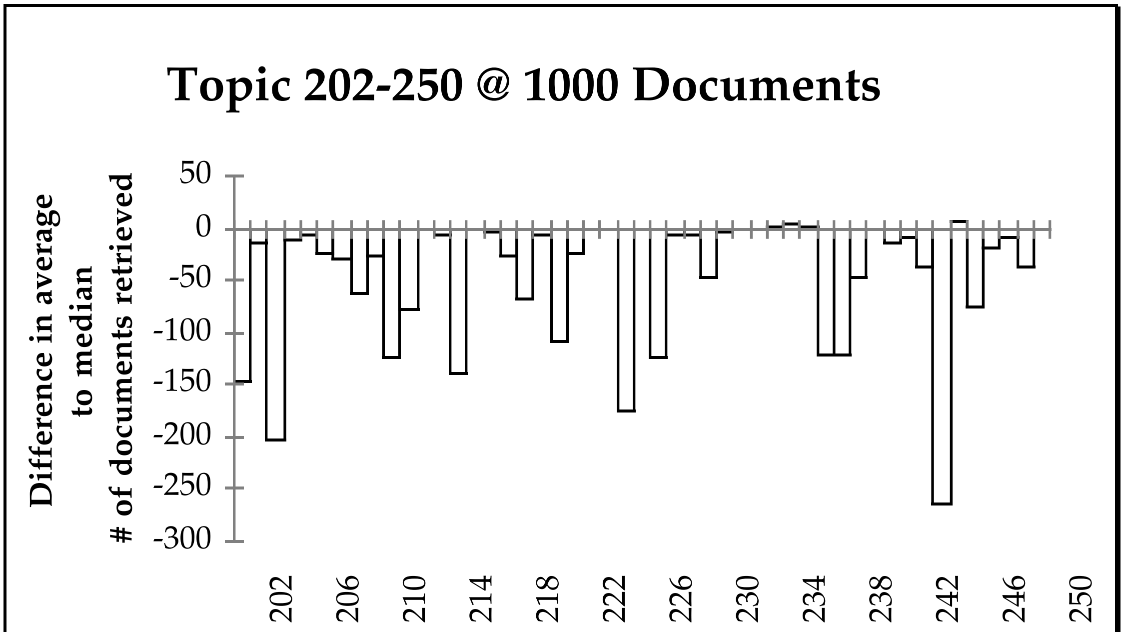


Figure 5

Figure 6 below summarizes the differences in R-precision for fsclt2 with respect to the TREC-4 median; these data are nearly identical to those in Figure 3.

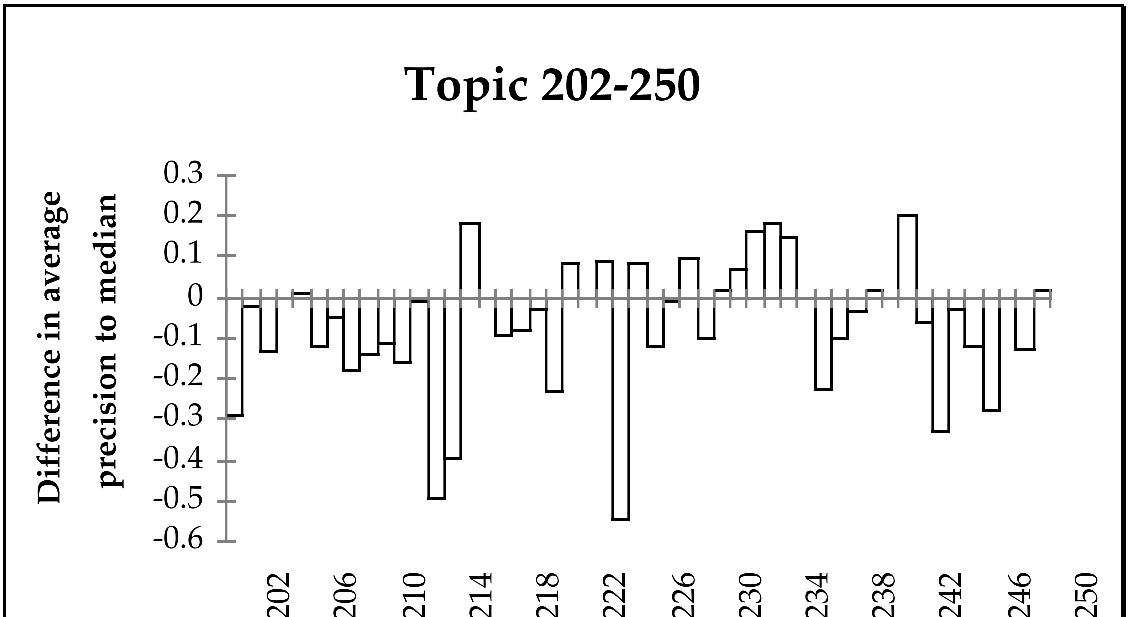


Figure 6

5.2.2 Success/Failure Analysis for fsclt2

The overall average precision and R-precision scores for fsclt2 were worse than those for fsclt 1, although topics 224, 226, 240 and 248 showed slightly-improved R-precision scores at 100 documents. In 12 cases, the average precision was improved. In 2 cases, more relevant documents appeared in the top 20 documents than had in fsclt1. Given the unsatisfactory nature of the results, we chose to ignore these few successes and concentrate on failure analysis for fsclt2 data. We began with a topic-by-topic comparison of results from the two TREC-4 experiments. By examining differences between 100 and 1000 document scores within and between experiments for each topic, 5 topics were identified for which R-precision scores dropped by 2 or 3 levels (as listed in Tables 1 and 2). Topics 240 and 248 were examined in detail.

For topic 240, the original query formulation was *terroris* AND (control OR prevent OR prevention OR combat).* The title of the first (i.e., most relevant) document was *AP-Carter gratified by new wave of popularity….* Candidate terms available for automated query expansion were: *situations, daydream, 0950, 0921, effusive 1433edt, deporting, complimented, am-carter, ax2*.. The actual expansion terms used, based on the selection parameters, were *am-carter* and *ax2*.. Clearly, the terms used for query expansion were irrelevant to the topic and could not have a positive effect on set re-ordering.

For topic 248, the original query formulation was *(blind OR blindness) AND (electronic OR technology)*. The title of the first (most relevant) document was : *Pentagon joins battle....* This document was a multi-section document;

an article on river blindness appeared after the initial Pentagon piece. The ten most frequent terms were : *0326est, bissau, government-industry, burkina, faso, benin, togo, deplete, niger, ozone-damaging.* Based on the term selection algorithm, the only term chosen for query expansion was *ozone-damaging.* It is clear that the term selection algorithm was not appropriate, given a document of this type.

Review of individual topic data led to the inescapable conclusion that the relevance feedback algorithm was flawed. Upon further examination, a bug was found in the custom client created for the TREC-4 experiment. Rather than selecting the three most relevant documents for use in the automated query expansion, the system selected only the first document, employing the extracted terms three times. As would be expected, the bug's effect seriously compromised our ability to judge the usefulness of the query expansion tool and rendered the fsclt2 results only minimally useful.

6 Discussion of FS Consulting TREC-4 Results

The MPS search engine is designed to operate in an interactive setting, where quick response and high precision are generally preferable to high recall. (High recall can be achieved by creating several different queries for the same topic; this is the recommended search strategy when high recall searches are required). The TREC-4 results for fsclt1 and fsclt2 show evidence of this design decision, as most queries, even well-constructed ones, returned less than 1000 documents. Also, levels of precision/recall were generally good at 100 documents, but poor when one looked at the top 1000 document. Results indicated a need for improvement in both searcher and system domains.

6.1 Searcher improvements

Examination of the query formulations indicated that the searcher did not take full advantage of system features, and constructed queries that were not optimal for the search system. For example, multi-concept nested query statements reduced recall in many of the searches. Whether or not these formulations are typical of average searchers, it seems clear that the training period did not produce sufficient understanding of the system's strengths and weaknesses, nor were optimal query models presented and reinforced. The ad hoc track's constraint against interactive involvement of the searcher contributed to the problem, since the user could not see results of one search and modify her strategies accordingly.

6.2 System improvements

As indicated earlier, several system problems contributed to poor test results.

Flaws in the custom client application caused the automated query expansion tool to work improperly. Aside from flaws in the custom client application, the relevance feedback algorithm (i.e., the selection criteria for picking expansion terms) did not work as desired. In some cases, it worsened precision/recall scores between experiments 1 and 2. Indexing the entire document, including news message header information, introduced inappropriate search terms. This reduced efficacy of the relevance feedback algorithm and contributed to the poor results demonstrated in fsclt2.

6.3 Initial Post TREC-4 Improvements

After success and failure analysis on the TREC-4 results, we explored modifications at the searcher and system level. In both areas, slight modifications produced improved results, as outlined below.

6.3.1 Searcher modifications

Six of the original query formulations for TREC-4 topics were reformulated as indicated below, and run against the TREC-4 databases.

Topic 211: *((DWI OR drunk) AND (drive OR driving) AND (penalties)) AND mortality* became *DWI OR (drunk AND (drive OR driving) AND penalties)*
Outcome: 31 relevant documents retrieved, compared to 0 in fsclt1.

Topic 214: *"Self-induced hypnosis"* became *hypnosis*
Outcome: 5 relevant documents retrieved compared to 0 in fsclt1.

Topic 215: *"infant mortality rate" AND ("United States" OR U.S. OR America) AND (higher or lower)* became (alt. 1) *"infant mortality" AND (higher or lower)* and (alt. 2) *"infant mortality"*
Outcome: Alt.1 found 61 relevant documents and Alt.2 145 relevant documents, compared to 20 in fsclt1.

Topic 243: *"fossil fuel" AND (utilization or use)* became *"fossil fuel"*
Outcome: 36 relevant documents retrieved, compared to 0 in fsclt1.

Topic 244*: ("United States" OR U.S. OR America) AND Japan AND "trade balance"* became *Japan AND (trade or balance)*
Outcome: 267 relevant documents retrieved, compared to 35 in fsclt1.

Topic 249: *weather AND "rain forest" AND (deplete or destroy OR disappear)* became (Alt.1) *"rain forest" AND (deplete OR destroy OR disappear*); (Alt.2)*"rain forest" AND deplet* OR destroy OR disappear)* ; (Alt.3) *"rain forest" OR rainforest AND destr*;* (Alt.4) *"rain forest" OR rainforest OR rainforest*

Outcome: Alt.1 retrieved 5 relevant documents, Alt.2 retrieved 7, Alt.3 retrieved 22 and Alt.4 retrieved 35 relevant documents compared to 1 in fsclt1.

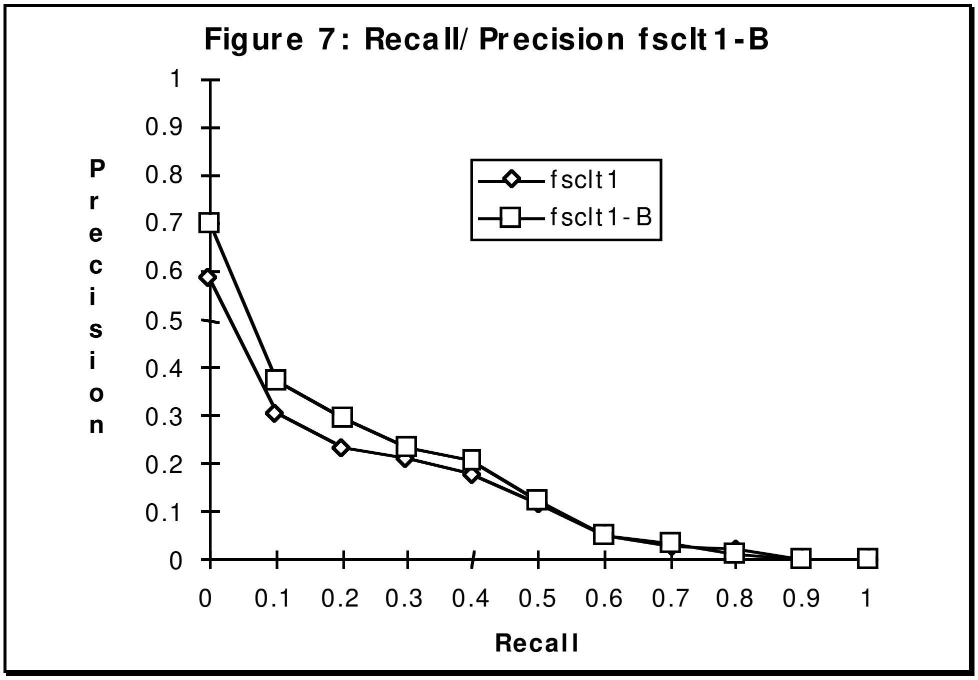
As the outcomes indicate, simplifying query formulations significantly improved retrieval results. For future experiments, training will be tuned to demonstrating outcomes of different query constructions.

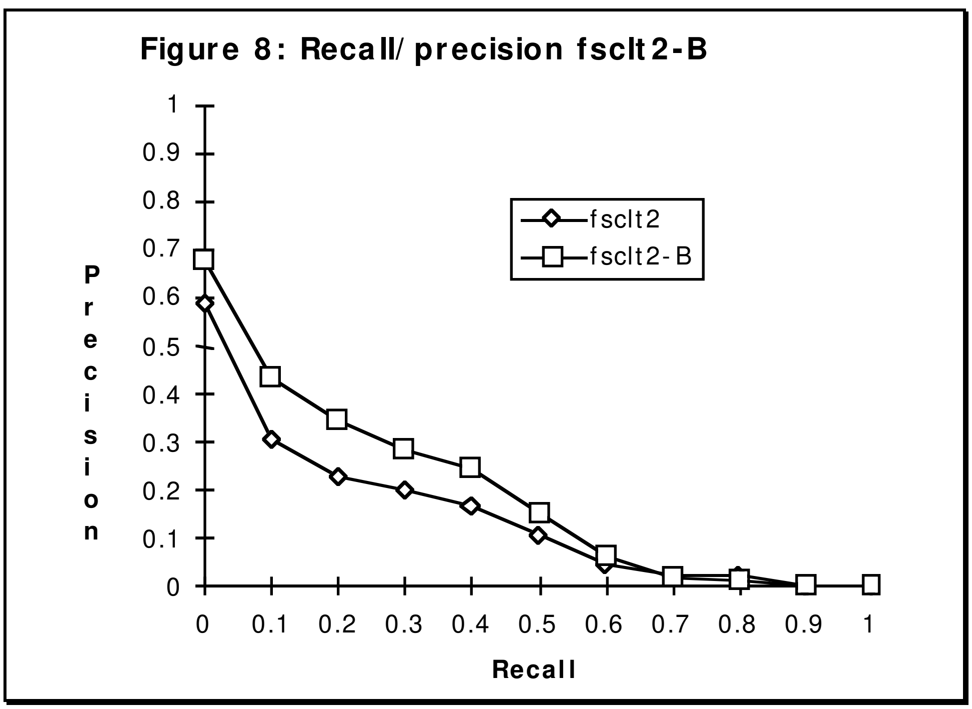
6.3.2 System modifications

Several aspects of the search mechanisms and relevance feedback algorithms were reworked. Using the original query formulations for TREC-4 ad hoc topics, the same two experiments were performed. Overall results for the post-TREC experiments:

```
Queryid (Num):      fsclt1-B
Total number of documents over all queries
    Retrieved:    11060
    Relevant:      6126
    Rel_ret:       1181
Interpolated Recall - Precision Averages:
    at 0.00       0.7018
    at 0.10       0.3695
    at 0.20       0.2946
    at 0.30       0.2329
    at 0.40       0.2020
    at 0.50       0.1189
    at 0.60       0.0472
    at 0.70       0.0308
    at 0.80       0.0207
    at 0.90       0.0113
    at 1.00       0.0000
Average precision (non-interpolated) over all rel docs
                  0.1532
Precision:
  At    5 docs:   0.4583
  At   10 docs:   0.4146
  At   15 docs:   0.3819
  At   20 docs:   0.3625
  At   30 docs:   0.3194
  At  100 docs:   0.1633
  At  200 docs:   0.0981
  At  500 docs:   0.0464
  At 1000 docs:   0.0246
R-Precision (precision after R (= num_rel for a query) docs
retrieved):
Exact:            0.2028
```

Figures 7 and 8 below compare the recall-precision results for the modified systems with results from the official fsclt1 and fsclt2 experiments.





The original results (fsclt1 and fsclt2) track each other very closely. The new test results (fsclt1-B and fsclt2-B) show improved precision in both experiments.

7 Future Work

The TREC-4 experiments provided baseline results and in a non-interactive environment and allowed exploration of possible directions for future work. Several themes emerged that will guide our research efforts in preparation for participation in TREC-5.

- The system will be tuned and improved. We will explore the effects of searching only selected fields in full-text documents. The query expansion tool will continue to be tested and revised. Additional relevance feedback algorithms will also be tested. A user interface will be constructed to allow the searcher to review results and make other decisions about search parameters.

- Additional examples of manual query formulations will be gathered and tested for TREC-4 topics, in order to build and improve the model of an 'average' searcher. Efforts will be made to gather formulations from searchers with different backgrounds (e.g., librarians, medical students, academic faculty, administrative and clerical staff). These data will be used to improve searcher training, and to suggest additional user tools.

- An interactive experiment will be designed based on improvements and searcher modeling to be undertaken this year.