Overview of the Third Text REtrieval Conference (TREC-3)

Donna Harman

National Institute of Standards and Technology
Gaithersburg, MD. 20899

1. Introduction

In November of 1992 the first Text REtrieval Conference (TREC-1) was held at NIST [Harman 1993]. The conference, co-sponsored by ARPA and NIST, brought together information retrieval researchers to discuss their system results on a new large test collection (the TIPSTER collection). This conference became the first in a series of ongoing conferences dedicated to encouraging research in retrieval from large-scale test collections, and to encouraging increased interaction among research groups in industry and academia. From the beginning there has been an almost equal number of universities and companies participating, with an emphasis on exploring many different types of approaches to the text retrieval problem.

The research done by the participating groups in the three TREC conferences has varied, but has followed a general pattern. TREC-1 required significant system rebuilding by most groups due to the huge increase in the size of the document collection (from a traditional test collection of several megabytes in size to the 2 gigabyte TIPSTER collection). The TREC-1 results should therefore be viewed as very preliminary due to severe time constraints. The second TREC conference (TREC-2) occurred in August of 1993, less than 10 months after the first conference [Harman 1994a]. Many of the original TREC-1 groups were able to "complete" their system rebuilding and tuning, and in general the TREC-2 results show significant improvements over the TREC-1 results. In some senses, however, the TREC-2 results should be viewed as a baseline for more complex experimentation.

The TREC-3 results reflect some of that more complex experimentation. For some groups that meant more extensive experiments based on their basic system techniques. For other groups it involved trying techniques from other groups and exploring more hybrid approaches. Some groups tried approaches that were radically different from their original approaches. As should be expected, those groups new to TREC had the same scaling problems as seen in TREC-1.

This paper provides an overview of the TREC-3 conference, including a review of the TREC task, a very brief description of the test collection being used, and an overview of the results. The papers from the individual groups should be referred to for more details on specific system approaches.

2. The Task and the Participants

The three TREC conferences have all centered around two tasks based on traditional information retrieval modes: a "routing" task and an "adhoc" task. In the routing task it is assumed that the same questions are always being asked, but that new data is being searched. This task is similar to that done by news clipping services or by library profiling systems. In the adhoc task it is assumed that new questions are being asked against a static set of data. This task is similar to how a researcher might use a library, where the collection is known, but it is unknown what questions are likely to be asked.

A schematic of those tasks is shown in Figure 1.

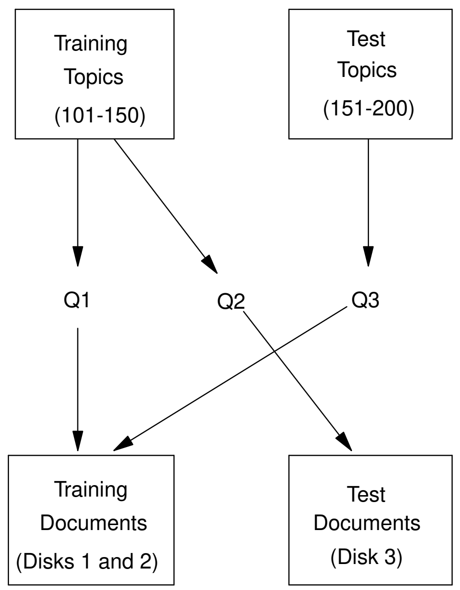

Figure 1. The TREC Task.

Table 1: TREC-3 Participants (14 companies, 19 universities)

Australian National University	Bellcore
Carnegie Mellon University/CLARITECH	CITRI, Australia
City University, London	Cornell University
Dublin City University	Environment Research Institute of Michigan
Fulcrum	George Mason University
Logicon Operating Systems	Mayo Clinic/Foundation
Mead Data Central	National Security Agency
New York University	NEC Corporation
Queens College	Rutgers University (two groups)
Siemens Corporate Research Inc.	Swiss Federal Institute of Technology (ETH)
TRW/Paracel	Universitaet Dortmund, Germany
University of California - Berkeley	University of Central Florida
University of Massachusetts at Amherst	VPI&SU (Virginia Tech)
University of Minnesota	University of Toronto
Universite de Neuchatel, Switzerland	Verity Inc.
West Publishing Co.	Xerox Palo Alto Research Center

In TREC the routing task is represented by using known topics and known relevant documents for those topics, but new data for testing. This is shown on the left side of Figure 1. The participants are given a set of known (or training) topics, shown in the top left-hand box, along with a set of known relevant documents (relevance judgments) for those topics. The topics consist of natural language text describing a user's information need (see section 3.3 for more description of the topics). These topics are used to create a set of queries (the actual input to the retrieval system) which are then used against the training documents. This is represented by Q1 in the diagram. Many sets of Q1 queries might be built to help adjust systems to this task, to create better weighting algorithms, and in general to train the system for testing. The results of this research are used to create Q2, the final routing queries to be used against the test documents.

The adhoc task is represented by using known documents, but new topics with no known relevant documents. This is shown on the right-hand side of Figure 1, where the 50 new test topics are used to create Q3 as the adhoc queries for searching against the training documents. The results from searches using Q2 and Q3 are the official test results sent to NIST.

In addition to clearly defining the tasks, other guidelines are used in TREC. These guidelines deal with the methods of indexing/knowledge base construction and with the methods of generating the queries from the supplied topics. In general, they are constructed to reflect an actual operational environment, and to allow as fair as possible separation among the diverse query construction approaches. Three generic categories of query construction were defined in TREC-3, based on the amount and kind of manual intervention used.

1. AUTOMATIC (completely automatic query construction)
2. MANUAL (manual query construction)
3. INTERACTIVE (use of interactive techniques to construct the queries)

There were 33 groups participating in TREC-3 (see Table 1), using a wide variety of retrieval techniques. One of the participants (Fulcrum) withdrew their results before the conference and therefore no results from this company appear in the proceedings. The participants were able to choose from three levels of participation: Category A, full participation, Category B, full participation using a reduced dataset (1/4 of the full document set), and Category C for evaluation only (to allow commercial systems to protect proprietary algorithms). Each participating group was provided the data and asked to turn in either one or two sets of results for each topic. When two sets of results were sent, they could be made using different methods of creating queries (AUTOMATIC, MANUAL, or INTERACTIVE), or by using different parameter settings for one query creation method. Groups could choose to do the routing task, the adhoc task, or both, and were requested to submit the top 1000 documents retrieved for each topic for evaluation.

TREC-3 introduced a second language (Spanish) to the task, with four groups working with a small Spanish collection in addition to their work in English. This collection, and the results, are discussed in section 5.4.

3. The Test Collection (English)

3.1 Introduction

Like most traditional retrieval collections, there are three

Table 2: Document Statistics

Subset of collection	WSJ (disks 1 and 2) SJMN (disk 3)	AP	ZIFF	FR (disks 1 and 2) PAT (disk 3)	DOE
Size of collection (megabytes)					
(disk 1)	270	259	245	262	186
(disk 2)	247	241	178	211	
(disk 3)	290	242	349	245	
Number of records					
(disk 1)	98,732	84,678	75,180	25,960	226,087
(disk 2)	74,520	79,919	56,920	19,860	
(disk 3)	90,257	78,321	161,021	6,711	
Median number of terms per record					
(disk 1)	182	353	181	313	82
(disk 2)	218	346	167	315	
(disk 3)	279	358	119	2896	
Average number of terms per record					
(disk 1)	329	375	412	1017	89
(disk 2)	377	370	394	1073	
(disk 3)	337	379	263	3543	

distinct parts to this collection -- the documents, the questions or topics, and the relevance judgments or "right answers." These test collection components are discussed very briefly in the rest of this section. For a more complete description of the collection, see [Harman 1994b].

3.2 The Documents

The documents were distributed as CD-ROMs with about 1 gigabyte of data each, compressed to fit. The following shows the actual contents of each disk.

Disk 1

- WSJ -- *Wall Street Journal* (1987, 1988, 1989)
- AP -- *AP Newswire* (1989)
- ZIFF -- Articles from *Computer Select* disks (Ziff-Davis Publishing)
- FR -- *Federal Register* (1989)
- DOE -- Short abstracts from DOE publications

Disk 2

- WSJ -- *Wall Street Journal* (1990, 1991, 1992)
- AP -- *AP Newswire* (1988)
- ZIFF -- Articles from *Computer Select* disks
- FR -- *Federal Register* (1988)

Disk 3

- SJMN -- *San Jose Mercury News* (1991)
- AP -- *AP Newswire* (1990)
- ZIFF -- Articles from *Computer Select* disks
- PAT -- U.S. Patents (1993)

Table 2 shows some basic document collection statistics. Note that although the collection sizes are roughly equivalent in megabytes, there is a range of document lengths across collections, from very short documents (DOE) to very long (FR). Also the range of document lengths within a collection varies. For example, the documents from AP are similar in length (the median and the average length are very close), but the WSJ, ZIFF and especially FR documents have much wider range of lengths within their collections.

The documents are uniformly formatted into SGML, with a DTD included for each collection to allow easy parsing.

<DOC>
<DOCNO> WSJ880406-0090 </DOCNO>
<HL> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </HL>
<AUTHOR> Janet Guyon (WSJ Staff) </AUTHOR>
<DATELINE> NEW YORK </DATELINE>
<TEXT>

American Telephone & Telegraph Co. introduced the first of a new generation of phone services with broad

.

.

</TEXT>
</DOC>

3.3 The Topics

In designing the TREC task, there was a conscious decision made to provide "user need" statements rather than more traditional queries. Two major issues were involved in this decision. First there was a desire to allow a wide range of query construction methods by keeping the topic (the need statement) distinct from the query (the actual text submitted to the system). The second issue was the ability to increase the amount of information available about each topic, in particular to include with each topic a clear statement of what criteria make a document relevant.

The new topics used in TREC-3 reflect a slight change in this decision. The topics in TREC-1 and 2 (topics 1-150) were not only very long, but contained complex structures. These topics were designed to mimic a real user's need, and were written by people who are actual users of a retrieval system. However they were intended to represent long-standing information needs for which a user might be willing to create elaborate topics, and therefore are more suited to the routing task than to the adhoc task, where users are likely to ask much shorter questions.

The new topics used in TREC-3 (topics 151-200) are not only much shorter, but missing the complex structure of the earlier topics. In particular the concepts field has been removed. This field contained a mini-knowledge base about a topic such as a real searcher might possess. The field was removed because it was felt that real adhoc questions would not contain this field, and because inclusion of the field discouraged research into techniques for expansion of "too short" user need expressions. Note that the shorter topics do not create a problem for the routing task, as experience in TREC-1 and 2 has shown that the use of the training documents allows a shorter topic (or no topic at all).

In addition to being shorter, the new topics were written by the same group of users that did the assessments. Specifically, each of the new topics (numbers 151-200) were developed from a genuine need for information brought in by the assessors. Each assessor constructed his/her own topics from some initial statements of interest, and performed all the relevance assessments on these topics (with a few exceptions).

The following is one of the new topics used in TREC-3. Each topic is formatted in the same standard method to allow easier automatic construction of queries.

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

<narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.

</top>

3.4 The Relevance Judgments

The relevance judgments are of critical importance to a test collection. For each topic it is necessary to compile a list of relevant documents; hopefully as comprehensive a list as possible. All three TRECs have used the pooling method [Sparck Jones & van Rijsbergen 1975] to assemble the relevance assessments. In this method a pool of possible relevant documents is created by taking a sample of documents selected by the various participating systems. This sample is then shown to the human assessors. The particular sampling method used in TREC is to take the top X documents retrieved by each system for a given topic and merge them into the pool for assessment. This is a valid sampling technique since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first.

Evaluation of retrieval results using the assessments from this sampling method is based on the assumption that the vast majority of relevant documents have been found and that documents that have not been judged can be assumed to be not relevant. A test of this assumption was made between TREC-2 and TREC-3, using TREC-2 results. Thirty-six (18 adhoc and 18 routing) topics were selected for additional relevance assessments, using a pseudo-random selection based only on the number of original relevant documents and on selecting equal numbers of topics from each assessor. For each selected topic, a new pool of documents was created by taking the top 200 documents from seven different runs known to achieve good results and to have little overlap in their document selection. New judgments were made on this pool, using the same judges that made the original decisions for each topic.

Table 3 gives the results of this test. On average, 30 new relevant documents (16%) were found for each of the top-

Table 3: Analysis of Completeness of Relevance Judgments (TREC-2)

Percent New Rel.	No. of Topics	Average New Rel.	Average Total Rel.	Average No. Jud.	Average "Hardness"
0%	5	0	46	381	0.3477
1-9%	11	10	173	257	0.4190
10-19%	9	36	277	343	0.2610
20-29%	6	47	185	190	0.3660
30-33%	5	73	242	233	0.5212
Average (over all 36 topics)		30	193	282	
Median		21	190	220	
Average (over the 18 routing topics)		18	188	373	
Median		8	160	376	
Average (over the 18 adhoc topics)		42	197	190	
Median		28	209	150	

Table 4: Overlap of Submitted Results

	Adhoc			Routing		
	Possible	Actual	Relevant	Possible	Actual	Relevant
TREC-1	3300	1279 (39%)	277 (22%)	2200	1067 (49%)	371 (35%)
TREC-2	4000	1106 (28%)	210 (19%)	4000	1466 (37%)	210 (14%)
TREC-3						
at 100	4800	1005 (21%)	146 (15%)	4900	703 (14%)	146 (21%)
at 200	9600	1946 (20%)	196 (10%)	9800	1333 (14%)	187 (14%)

ics, with a median of only 21 (11%) new relevant documents per topic. The median is much lower than the average because of the relatively large number of new documents found for those five topics with over 30% additional relevant documents found.

Table 3 also shows that there is some correlation between the number of new relevant documents found and the original number of relevant documents, particularly in that topics with few relevant documents initially tended to have few new ones found. In contrast, there is no correlation between the number of new relevant documents found and the number of new judgments made, or between the number of new relevant found for a topic and the "hardness" of that topic (a measure of average system performance for that topic). More new relevant documents were found for the adhoc task than for the routing task. This may reflect more "available" relevant documents for the adhoc task (twice the amount of searchable text) or may be caused by the more complete and accurate queries used in routing task due to the training data.

A different measure of the effect of pooling can be seen by examining the overlap of retrieved documents. Table 4 shows the statistics from the merging operations in the three TREC conferences. For TREC-1 and TREC-2 the top 100 documents from each run (33 runs in TREC-1 and 40 runs in TREC-2) could have produced a total of 3300 and 4000 documents to be judged (for the adhoc task). The average number of documents actually judged per topic (those that were unique) was 1279 (39%) for TREC-1 and 1106 (28%) for TREC-2. Note that even though the number of runs has increased by more than 20% (adhoc), the number of unique documents found has actually dropped. The percentage of relevant documents found, however, has not changed much. The more accurate results going from TREC-1 to TREC-2 mean that fewer "noisy" nonrelevant documents are being found by the systems. This trend continued in TREC-3, even though the pooling method was changed.

Because of expected constraints in assessor time, only one run from each TREC-3 group was judged, with the groups specifying which run. However, due to the increase in overlap (as shown in Table 4), and more efficient judging, extra time became available and the decision was made to judge the top 200 documents for those runs. Table 4 gives the results of the TREC-3 mergings at both 100 documents and 200 documents. The percentage of unique documents found continues to drop compared with TREC-2, with a major drop for the routing. The total number of relevant documents found in TREC-1, TREC-2, and TREC-3 has dropped only somewhat, however, and that drop has been caused by a deliberate tightening of the topics between TREC-1 and TREC-2. Table 4 also shows

Table 5: Analysis of Pooling Methodologies (Adhoc)

TREC-2 -- Relevant Documents Found in "Second" Run			
Percent New Rel.	No. of Topics	Average New Rel.	Average No. Rel.
0%	0	-	-
1-9%	6	9	123
10-19%	19	26	163
20-29%	19	68	274
30-36%	5	109	296
Average		48	210
Median		30	201
TREC-3 -- Relevant Documents Found above 100			
Percent New Rel.	No. of Topics	Average New Rel.	Average No. Rel.
0%	1	0	85
1-9%	12	3	65
10-19%	7	13	96
20-29%	22	59	237
30-36%	8	137	381
Average		50	196
Median		30	122

Table 6: Analysis of Pooling Methodologies (Routing)

TREC-2 -- Relevant Documents Found in "Second" Run			
Percent New Rel.	No. of Topics	Average New Rel.	Average No. Rel.
0%	4	0	6
1-9%	8	4	61
10-19%	21	33	220
20-29%	11	88	345
30-36%	6	84	259
Average		44	210
Median		33	163
TREC-3 -- Relevant Documents Found above 100			
Percent New Rel.	No. of Topics	Average New Rel.	Average No. Rel.
0%	7	0	24
1-9%	9	6	106
10-19%	16	19	129
20-29%	16	94	354
30-36%	2	91	249
Average		41	187
Median		13	123

the drop in relevant documents found beyond the 100 document cutoff. This not only reflects the ranking done by the systems, but shows the diminishing numbers of relevant documents to be found even as the judged pool continues to grow.

The use of a different pooling method in TREC-3 provided a chance to compare the two methods. Tables 5 and 6 show this comparison. The first method (that used in TREC-2) took the top 100 documents from two runs, whereas the second method (that used in TREC-3) took the top 200 documents from a single run. The "base" for both methods is the top 100 documents in the single or "first" run. The additional documents to be compared are the number of relevant documents in the top 100 for the "second" run (TREC-2) versus the number of relevant documents in the second 100 in the single run for TREC-3.

Table 5 shows that both pooling methods worked equally well for the adhoc task. About the same numbers of relevant documents were found by each method, with similar averages, medians, and distributions of "new" relevant documents across the topics. This verifies the TREC-2 completeness experiments shown in Table 3, in that the average and median number of "new" documents found beyond the 100 document cutoff is similar to those found in TREC-3.

For the routing task, however, Table 6 shows that the first pooling method (TREC-2) seems to have found more relevant documents (higher median). Whereas this could reflect something about the different topics used in TREC-2 and TREC-3, it is more likely a reflection of the difference between system performance in the adhoc and routing tasks. Routing runs are generally more accurate in finding documents and more effective in ranking them, due to the availability of training data. Therefore the second 100 documents are less likely to contain additional relevant documents for the routing task than for the adhoc task. Again this verifies the completeness experiments shown in Table 3, which show far fewer new relevant documents being found for the routing task after the 100 document cutoff.

This analysis suggests a return to the TREC-2 pooling methodology, and that is what is planned for TREC-4. Participating groups would also prefer judgments on both official runs as this allows more exactness in evaluating run variations.

After pooling, each topic was judged by a single assessor to insure the best consistency of judgment. Some testing of this consistency was done after TREC-2, and showed an average agreement between two judges of about 80%. More consistency testing will be done in the future.

4. Evaluation

An important element of TREC is to provide a common evaluation forum. Standard recall/precision and recall/fallout figures have been calculated for each TREC system and are shown in Appendix A, along with some single evaluation measures for each system. A detailed explanation of the measures is also included in the appendix. New for TREC-3 is a histogram for each system showing performance on each topic. In general more emphasis is being placed on a per topic analysis this year in an effort to get beyond the averages. (Although work has been done to find statistical differences between the averages, see paper "A Statistical Analysis of the TREC-3 Data" by Jean Tague-Sutcliffe and James Blustein.)

Additional data about each system was collected that describes system features and system timing, and allows some primitive comparison of the amount of effort needed to produce the results. The individual system descriptions are given in Appendix B.

5. Results

5.1 Introduction

One of the important goals of the TREC conferences is that the participating groups freely devise their own experiments within the TREC task. For some groups this means doing the routing and/or adhoc task with the goal of achieving high retrieval effectiveness performance. For other groups, however, the goals are more diverse and may mean experiments in efficiency, unusual ways of using the data, or experiments in how "users" would view the TREC paradigm.

The overview of the results discusses the effectiveness of the systems and analyzes some of the similarities and differences in the approaches that were taken. Additionally it points to some of the other experiments run in TREC-3 where results cannot be measured completely using recall/precision measures.

In all cases, readers are referred to the system papers in this proceedings for more details.

5.2 Adhoc Results

The adhoc evaluation used the new topics (topics 151-200) against the two disks of training documents (disks 1 and 2). A dominant feature of the adhoc task in TREC-3 was the removal of the concepts field in the topics (see more on this in the discussion of the topics, section 3.3) Many of the participating groups designed their experiments around techniques to expand the shorter and less "rich" topics.

There were 48 sets of results for adhoc evaluation in TREC-3, with 42 of them based on runs for the full data set. Of these, 28 used automatic construction of queries, 12 used manual construction, and 2 used interactive construction.

Figure 2 shows the recall/precision curves for the 6 TREC-3 groups with the highest non-interpolated average precision using automatic construction of queries. The runs are ranked by the average precision and only one run is shown per group (both official Cornell runs would have qualified for this set).

A short summary of the techniques used in these runs shows the breadth of the approaches. For more details on the various runs and procedures, please see the appropriate papers in this proceedings.

citya1 -- City University, London (see paper "Okapi at TREC-3" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used a probabilistic term weighting scheme similar to that used in TREC-2, but expanded the topics by up to 40 terms (average around 20) automatically selected from the top 30 documents retrieved. They also used dynamic passage retrieval in addition to the whole document retrieval in their final ranking.

INQ101 -- University of Massachusetts at Amherst (see paper "Document Retrieval and Routing Using the INQUERY System" by John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar) used a version of probabilistic weighting that allows easy combining of evidence (an inference net). Their basic term weighting formula (and query processing) was simplified from that used in TREC-2, and they also used passage retrieval and whole document information in their ranking. The topics were expanded by 30 phrases that were automatically selected from a phrase "thesaurus" that had been previously built automatically from the entire corpus of documents.

CrnlEA -- Cornell University (see paper "Automatic Query Expansion Using SMART: TREC-3 by Chris Buckley, Gerard Salton, James Allan and Amit Singhal) used the vector-space SMART system, with term weighting similar to that done in TREC-2. The top 30 documents were used in a Rocchio relevance feedback technique to massively expand (500 terms + 10 phrases) the topics. No passage retrieval was done in this run; the second Cornell run (*CrnlLA*) used their local/global weighting schemes (with no topic expansion).

westp1 -- West Publishing Company (see paper "TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System" by Paul Thompson, Howard Turtle,

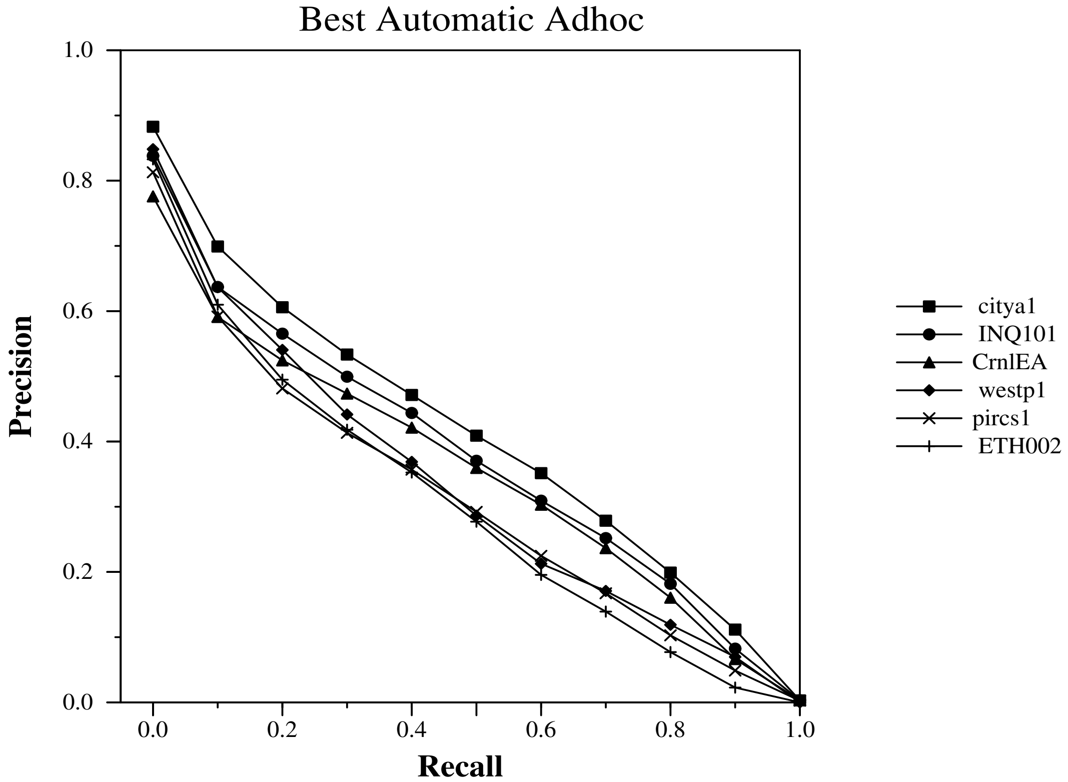

Figure 2. Best Automatic Adhoc Results.

Bokyung Yang and James Flood) used their commercial product (WIN) which is based on the same inference method used in *INQ101*. Both passages and whole documents were used in document ranking, but only minimal topic expansion was used, with that expansion based on preconstructed general-purpose synonym classes for abbreviations and other exact synonyms.

pircs1 -- Queens College, CUNY (see paper "TREC-2 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and D.D. Lewis) used a spreading activation model on subdocuments (550-word chunks). Topic expansion was done by allowing activation from the top 6 documents in addition to the terms in the original topic. The highest 30 terms were chosen, with an average of 11 of those not in the original topic.

ETH002 -- Swiss Federal Institute of Technology (ETH) (see paper "Improving a Basic Retrieval Method by Links and Passage Level Evidence" by Daniel Knaus, Elke Mittendorf and Peter Schäuble) used a completely new method in TREC-3 based on combining information from three very different retrieval techniques. The three techniques are a vector-space system, a passage retrieval method using a Hidden Markov model, and a "topic expansion" method based on document links generated automatically from analysis of common phrases.

The dominant new themes in the automatic adhoc runs are the use of some type of term expansion beyond the terms contained in the shorter (TREC-3) topics, and some form of passage or subdocument retrieval element. Note that term expansion is mostly a recall device; adding new terms to a topic increases the chances of matching the wide variation of terms usually found in relevant documents. But adding terms also increases the "noise" factor, so accuracy may need to be improved via a precision device, and hence the use of passages, subdocuments, or more local weighting.

Two main types of term expansion were used by these top groups: term expansion based on a pre-constructed thesaurus (for example the INQUERY PhraseFinder) and term expansion based on selected terms from the top X documents (as done by City, Cornell, and PIRCS). Both techniques worked well. The top 3 runs (*citya1*, *INQ101*, and *CrnlEA*) have excellent performance (see Figure 2) in the "middle" recall range (30 to 80%), with this performance likely coming from the query expansion.

The use of the top 30 documents as a source of terms, as opposed to using the entire corpus, should be sensitive to the quality of the documents in this initial set. Notably, for 6 of the 8 topics in which the *INQ101* run was superior (a 20% or more improvement in average precision) to the *citya1* run, the *INQ101* run was also superior to the *CrnlEA* run. These topics tended to have fewer relevant documents, but also tended to be topics for which the systems bringing terms in manually (such as by manually selecting from a thesaurus or outside sources) did well.

Clearly there are topics in which this technique does not work well, but it does seem to provide an excellent focussing effect for many topics. This may not be the case outside of TREC, where there are fewer relevant documents. However, this type of expansion should be considered a worthwhile tool for query modification, especially for environments where no thesaurus exists.

Another factor in topic expansion is the number of terms being added to the topics. The average number of terms in the queries is widely varied, with the City group averaging around 50 terms (20 terms from expansion), the INQUERY system using around 100 terms on average, and the Cornell system using 550 terms on average. This huge variation seemed to have little effect on results, largely because each group found the level of topic expansion appropriate for their retrieval techniques. The *citya1* run tended to "miss" more relevant documents than the *CrnlEA* run (7 topics were seriously hurt by this problem), but was better able to rank relevant documents within the 1000 document cutoff so that more relevant documents appeared in the top 100 documents. This better ranking could have happened because of the many fewer terms that were used, or could be caused by the use of passage retrieval in the City run.

The use of passages or subdocuments to reduce the noise effect of large documents has been used for several years in the PIRCS system. City, INQUERY and Cornell all did many experiments for TREC-3 to first determine the correct length of a passage, and then to find the appropriate use of passages in their ranking schemes. INQUERY and Cornell use overlapped passages of fixed length (200 words) as compared to City's non-overlapped passages of 4 to 30 paragraphs in length. All three systems use information from passages and whole documents retrieved rather than passage retrieval alone. (Cornell's version of this is called local/global weighting.) Both INQUERY and City combined the passage retrieval with query expansion; Cornell did two separate runs.

Note that the first two groups used passage retrieval to improve ranking and to regain the precision lost during the topic expansion. Cornell did not combine these operations even though they used term expansions on the order of 500 terms. The vector-space model seems less susceptable to "noise", as has been demonstrated in routing tasks. However in comparing the 2 Cornell runs, there were 16 topics in which the local/global run (*CrnlLA*) was superior, with 12 of these from better ranking, as opposed to only 8 topics that were superior in the expanded run (*CrnlEA*), 6 of which came from finding more relevant documents. A way of combining these runs should help performance, even for Cornell.

The *westp1* run did not use topic expansion, although a mixture of passages and whole documents was used in the final ranking of documents. The performance has suffered for this in the middle recall range. West Publishing used their production system to see how far it differed from the research systems and therefore did not want to use more radical topic expansion methods. Additionally they used a shortened topic (title + description + first sentence of narrative) because it was more similar in length to the topics submitted by their users. The *INQ101* run had 18 topics with superior performance to the *westp1* run, mostly because of new relevant documents being retrieved to the top 1000 document set. The 11 topics in which the *westp1* was superior to the *INQ101* run were mostly caused by better ranking for those topics.

The *pircs1* system used both passage retrieval (subdocuments) and topic expansion. This system used far fewer top documents for expansion (the top 6 as opposed to the top 30), and this may have hurt performance. There were 22 topics in which the *INQ101* run was superior to the *pircs2* run, and these were mostly because of missed relevant documents. Even though both systems added about the same number of expansion terms, using only the top 6 documents as a source of terms for spreading activation might have provided too much focussing of the concepts.

The *ETH001* run used both topic expansion and passages, in addition to a baseline vector-space system. Both the topic expansion and the passage determination were completely new (untried) techniques; additionally there are known difficulties in combining multiple methods. In comparison to the Cornell expansion results (*CrnlEA*), the main problems appear to be missed relevant documents for all 17 of the topics where the Cornell results were superior. The 8 topics with superior ETH results were mostly because of better ranking. Clearly this is a very promising approach and more experimentation is needed.

Table 7 shows a breakdown of improvements from expansion and passage retrieval that combines information from the non-official runs given in the individual papers. In general groups seem to be getting about 20% improvement over their own baselines (less for ETH and PIRCS), with that improvement coming in different percentages from passage retrieval or expansion, depending on the specific retrieval techniques being used.

Figure 3 shows the recall/precision curves for the 6 TREC-3 groups with the highest non-interpolated average precision using manual construction of queries. A short

Table 7: Comparison of Performance (Average Precision)
for Passage Retrieval and Topic Expansion

	base run	passages	expansion	both
City	0.337	-	0.388 (15%)	0.401 (19%)
INQUERY				
(11 pt. average)	0.318	0.368 (16%)	0.348 (9%)	0.381 (20%)
Cornell	0.2842	0.3302 (16%)	0.3419 (20%)	-
ETH	0.2578	0.2853 (11%)	0.2737 (6%)	0.2916 (13%)
PIRCS	-	0.2764	-	0.3001 (9%)

summary of the techniques used in these runs follows. Again, for more details on the various runs and procedures, see the appropriate papers in this proceedings.

INQ102 -- University of Massachusetts at Amherst. This run is a manual modification of the *INQ101* run, with strict rules for the modifications to only allow removal of words and phrases, modification of weights, and addition of proximity restrictions.

Brkly7 -- University of California, Berkeley (see paper "Experiments in the Probabilistic Retrieval of Full Text Documents" by William S. Cooper, Aitao Chen and Fredric C. Gey) is a modification of the *Brkly6* run, with that modification being the manual expansion of the queries by adding synonyms found from other sources. The *Brkly6* run uses a logistic regression model to combine information from 6 measures of document relevancy based on term matches and term distribution. The coefficients were learned from the training data in a manner similar to that done in TREC-2, but the specific set of measures used has been expanded and modified for TREC-3. No passage retrieval was done.

ASSCTV1 -- Mead Data Central, Inc (see paper "Query Expansion/Reduction and its Impact on Retrieval Effectiveness" by X. Allan Lu and Robert B Keefer) is also a manual expansion of queries using an associative thesaurus built from the TREC data. The retrieval system used in *ASSCTV1* is the SMART system.

VTc2s2 -- Virginia Tech (see paper "Combination of Multiple Searches" by Joseph A. Shaw and Edward A. Fox) used a combination of multiple types of queries, with 2 types of natural language vector-space queries and 3 types of manually constructed P-Norm (soft Boolean) queries.

pircs2 -- Queens College, CUNY. This run is a modification of the base PIRCS system to use manually constructed soft Boolean queries.

rutfua1 -- Rutgers University (see paper "Decision Level Data Fusion for Routing of Documents in the TREC3 Context: A Best Cases Analysis of Worst Case Results" by Paul B. Kantor) used data fusion methods to combine the retrieval ranks from three different retrieval schemes all using the INQUERY system. Two of the schemes used Boolean queries (one with ranking and one without) and the third used the same queries without operators.

The three dominant themes in the runs using manually constructed queries are manual modification of automatically generated queries (*INQ102*), manual expansion of queries (*Brkly7* and *ASSCTV1*) and combining of multiple retrieval techniques or queries. Three runs can be compared to a "baseline" run to check the effects of manual versus automatic query construction.

INQ102, the manually modified version of *INQ101*, had a 15% improvement in average precision over *INQ101*, and 17 topics that were superior in performance for the manual system (as opposed to only 3 for the automatic system). An analysis of those topics shows that many more relevant documents were in the top 1000 documents and the top 100 documents, probably caused by manually eliminating much of the noise that was producing higher ranks for nonrelevant documents. This noise elimination could have happened because many spurious terms had been manually removed from the queries (*INQ102* had an average of about 30 terms as opposed to nearly 100 terms in *INQ101*), or could have come from the use of the proximity operators.

The *Brkly7* run, a manually expanded version of *Brkly6*, used about the same number of terms as the *INQ102* run (around 36 terms on average), but the terms had been manually pulled from multiple sources (as opposed to editing an automatic expansion as done by INQUERY). The improvement from *Brkly6* to *Brkly7* is a 34% gain in average precision, with 25 topics having superior performance in the manually expanded run. Note however that there was no topic expansion done in the automatic *Brkly6* run, so this improvement represents the results of a good manual topic expansion over no expansion at all.

The INQUERY system outperforms the Berkeley system by 14% in average precision, with much of that difference coming in the high recall end of the graph (see Figure 3).

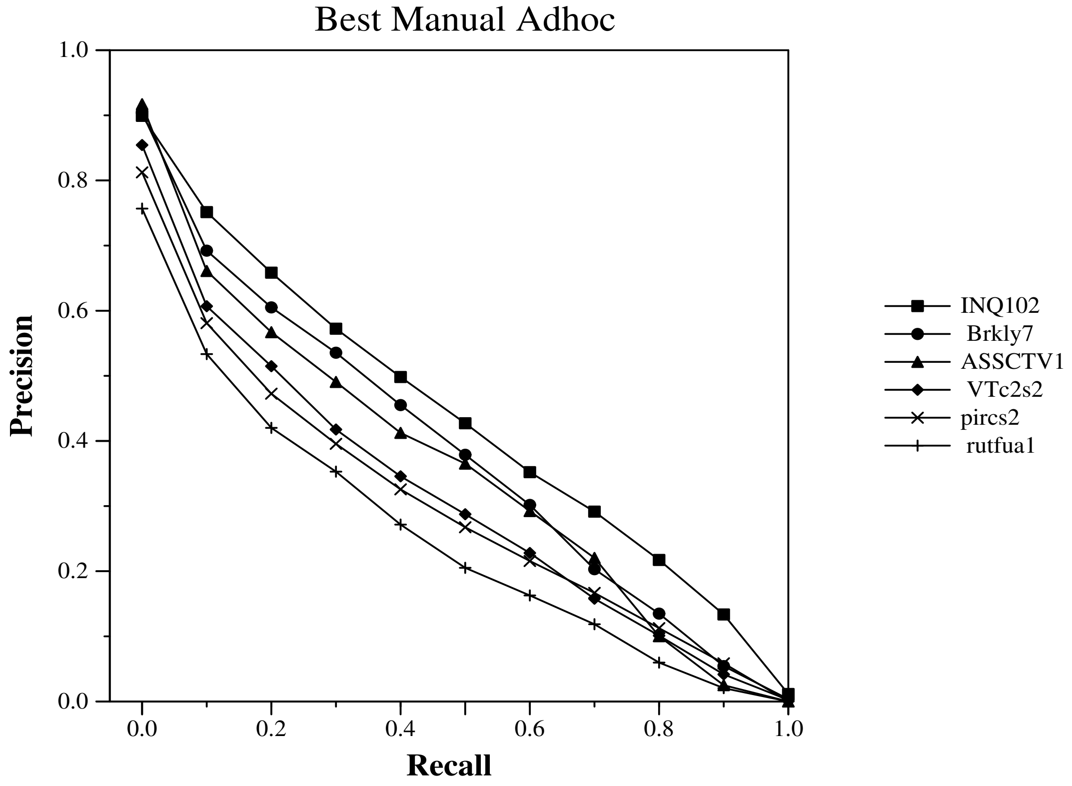

Figure 3. Best Manual Adhoc Results.

This is consistent with the difference in their topic expansion techniques in that the automatic expansion (even manually edited) is likely to bring in terms that users might not select from "non-focussed" sources.

The *ASSCTV1* run also represents a manual expansion effort, but using a pre-built thesaurus as opposed to using textual sources for the expansion. The topics were expanded to create a query averaging around 135 terms and then were run using the default Cornell SMART system. A comparison of the automatically expanded *CrnlEA* run and the manually expanded *ASSCTV1* run shows minimal difference in average precision, but superior performance in 18 of the topics for the manual expansion (as opposed to only 10 of the topics having superior performance for the automatic Cornell run). In both cases, the improvements come from finding more relevant documents because of the expansions, but different expansion methods help different topics.

The *pircs2* run is a manual query version of the baseline PIRCS system. A soft Boolean query is created from the topic, but no topic expansion is done. There is minimal difference in average precision between the two PIRCS runs, but more topics show superior performance for the soft Boolean query *pircs2* run (8 superior topics versus 4 superior topics for the topic expansion *pircs1* run). It is not clear whether this difference comes from the increased precision of the soft Boolean approach or from the relatively poor performance of the PIRCS term expansion results.

In TREC-3, as opposed to TRECs 1 and 2, the manual query construction methods perform better than their automatic counterparts. The removal of some of the topic structure (the concepts) has allowed differences to appear that could not be seen in earlier TRECs. Since topic expansion was necessary to produce top scores, the superiority of the manual expansion over no expansion in the Berkeley runs should not be surprising. Less clear is why the manual modifications in the *INQ102* run showed superior performance to the automatic run with no modifications. The likely explanation is that the automatic term expansion methods are relatively uncontrolled in TREC-3 and manual intervention plays an important role.

The last two groups in the top six systems using manual query construction used some form of combination of retrieval techniques. The Virginia Tech group (*VTc2s2*) combined the results of up to 5 different types of query construction (3 P-Norms with different P values and 2 vector-space, one short and one manually expanded) to create their results. They used a simple combination method (adding all the similarity values) and tested various combinations of query types. Their best result combined only two of the query types, one a P-Norm and one

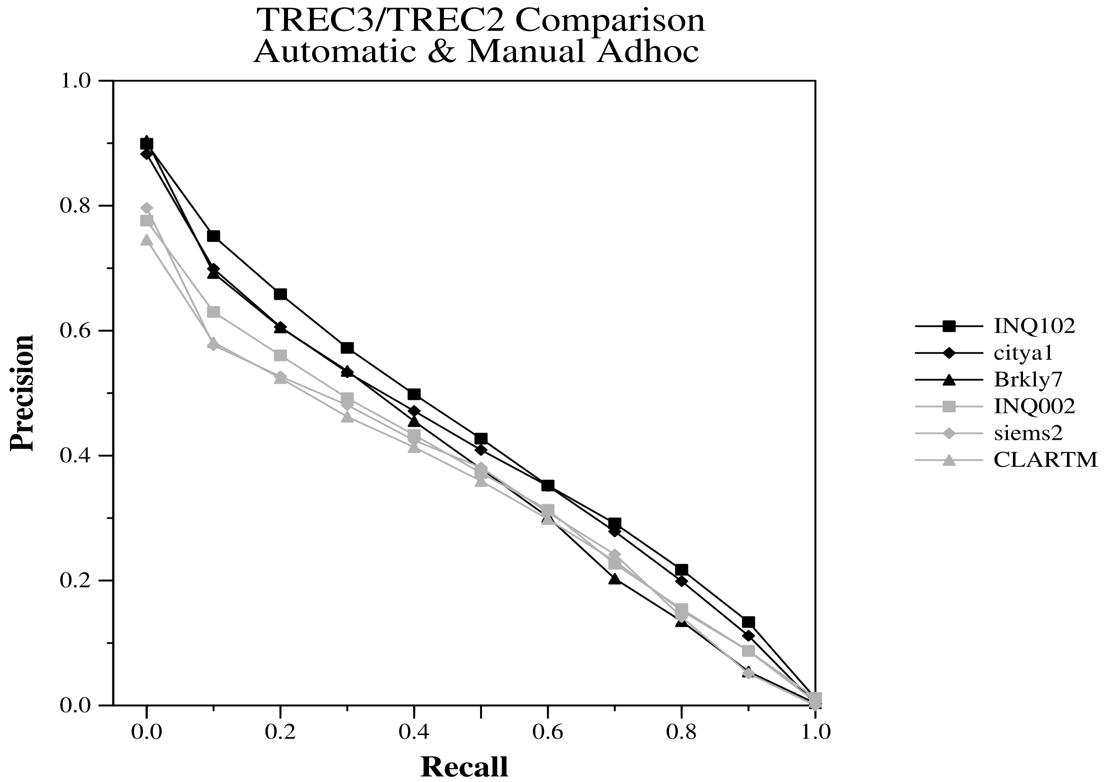

Figure 4. Comparison of Adhoc Results for TREC-2 and TREC-3

a vector-space. A series of additional runs (see paper for details) confirmed that the best method was to combine the results of the best two query techniques (the "long" vector-space and the P=2 P-Norm). They concluded that improvements from combining results only occurred when the input techniques were sufficiently different.

Although the Rutgers group (*rutfua1*) used more elaborate combining techniques, they came to the same conclusion. Combining different retrieval techniques offers improvements over a single technique (over 30% for the Virginia Tech group), but the input techniques need to be more varied to get further improvements. But the more varied the individual techniques, the more need for elaborate combining methods such as used in the *rutfua1* run. The automatic *ETH001* run best exemplifies the direction needed here; first getting "good" performance for three very different but complementary techniques and then discovering the best ways of combining results.

Several comments should be made with respect to the overall adhoc recall/precision averages. First, the better results are very similar and it is unlikely that there is any statistical difference between them. The Scheffe´ tests run by Jean Tague-Sutcliffe (see paper "A Statistical Analysis of the TREC-3 Data" by Jean Tague-Sutcliffe and James Blustein) show that the top 20 category A runs (manual and automatic mixed) are all statistically equivalent at the α=0.05 level. This lack of system differentiation comes from the very wide performance variation across topics (the cross-topic variance is much greater than the cross-system variance) and points to the need for more research into how to statistically characterize the TREC results.

As a second point, it should be noted that these adhoc results represent significant improvements over TREC-2. Figure 4 shows the top three systems in TREC-3 and the top three systems in TREC-2. This improvement was unexpected as the removal of the concepts section seemed likely to cause a considerable performance drop (up to 30% was predicted). Instead the advance of topic expansion techniques caused major improvements in performance with less "user" input (the concepts). Because of the different sets of topics involved, the exact amount of improvement cannot be computed. However the Cornell group has run older systems (those used in TREC-1 and TREC-2) against the TREC-3 topics. This shows an improvement of 20% for their expansion run (*CrnlEA*) over the TREC-2 system, and this is likely to be typical for many of the systems this year.

5.3 Routing Results

The routing evaluation used a subset of the training topics (topics 101-150 were used) against the disk of test documents (disk 3). Although this disk had been used in TREC-2, its use in TREC-3 was unexpected as new data had been promised. The last minute unavailability of this new data made the reuse of disk 3 necessary, but since groups had not been training with this disk (and no relevance judgments were available for this disk against topics 101-150), the routing results should not be biased by the reuse of old material.

The routing task in TREC has remained constant; however there has been a major evolution in the thrust of the research for this task. There was minimal training data for TREC-1, and most groups felt that their results were even more preliminary than for the adhoc results because the training data that was available was incomplete and inconsistent. This means that routing became a particularly interesting challenge in TREC-2 when adequate training data (the results from TREC-1 adhoc topics) became available.

The TREC-2 results therefore represent an excellent baseline of what could be achieved using traditional algorithms with large amounts of relevance information. Most notable was the effective use of the Rocchio feedback algorithm in SMART, where up to 500 new terms were added to the routing topics from the training data. Equally good results were achieved by a probabilistic system from the University of Dortmund, where only 30 terms were added, but very precise term weighting was learned from the training data. Manual construction of queries consistently gave poorer performance as the availability of training data allowed an automatic tuning of the queries that would be difficult to duplicate manually without extensive analysis.

For TREC-3, many groups made only minor modifications to their TREC-2 techniques (and concentrated on the adhoc task). There were a total of 49 sets of results for routing evaluation, with 46 of them based on runs for the full data set. Of the 46 systems using the full data set, 24 used automatic construction of queries, 18* used manual construction, and 4 used interactive query construction.

Figure 5 shows the recall/precision curves for the 12 TREC-3 groups with the highest non-interpolated average precision for the routing queries. The runs are ranked by the average precision and only one run per group is shown (both official runs sometimes would have qualified for this set). A short summary of the techniques used in these runs follows. For more details on the various runs and procedures, please see the appropriate papers in this proceedings.

cityr1 -- City University, London (see paper "Okapi at TREC-3" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used the same probabilistic techniques as for the adhoc task, but constructed the query using a very selective set of terms (17 on average) from the relevant documents.

pircs3 -- Queens College, CUNY (see paper "TREC-2 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and D.D. Lewis) used a spreading activation model based on the topic and on terms selected from about 35% of the relevant material.

INQ103 -- University of Massachusetts at Amherst (see paper "Document Retrieval and Routing Using the INQUERY System" by John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar) used the inference net engine (same as for the adhoc task), with topic expansion of about 60 terms selected from the relevant documents.

dortR1 -- University of Dortmund (see paper "Routing and Ad-hoc Retrieval with the TREC-3 Collection in a Distributed Loosely Federated Environment" by Nikolaus Walczuch, Norbert Fuhr, Michael Pollmann and Birgit Sievers) used the SMART retrieval system with a Rocchio relevance feedback expansion adding 12% new terms and 4% new phrases from the training documents.

lsir2 -- Bellcore (see paper "Latent Semantic Indexing (LSI): TREC-3 Report" by Susan Dumais) used the latent semantic indexing system to construct a reduced dimension vector centroid of the relevant documents (no use was made of the topics).

CrnlRR -- Cornell University (see paper "Automatic Query Expansion Using SMART: TREC-3 by Chris Buckley, Gerard Salton, James Allan and Amit Singhal) used the vector-space SMART system and a basic Rocchio relevance feedback algorithm adding about 300 terms and 30 phrases to the topic.

Brkly8 -- University of California, Berkeley (see paper "Experiments in the Probabilistic Retrieval of Full Text Documents" by William S. Cooper, Aitao Chen and Fredric C. Gey) used only the relevant documents to select a large number of terms (average 1,357 terms/topic) which were combined and weighted using a logodds formula. A chi-square test was used to select the terms.

---

* 11 of these runs were abbreviated runs from one group

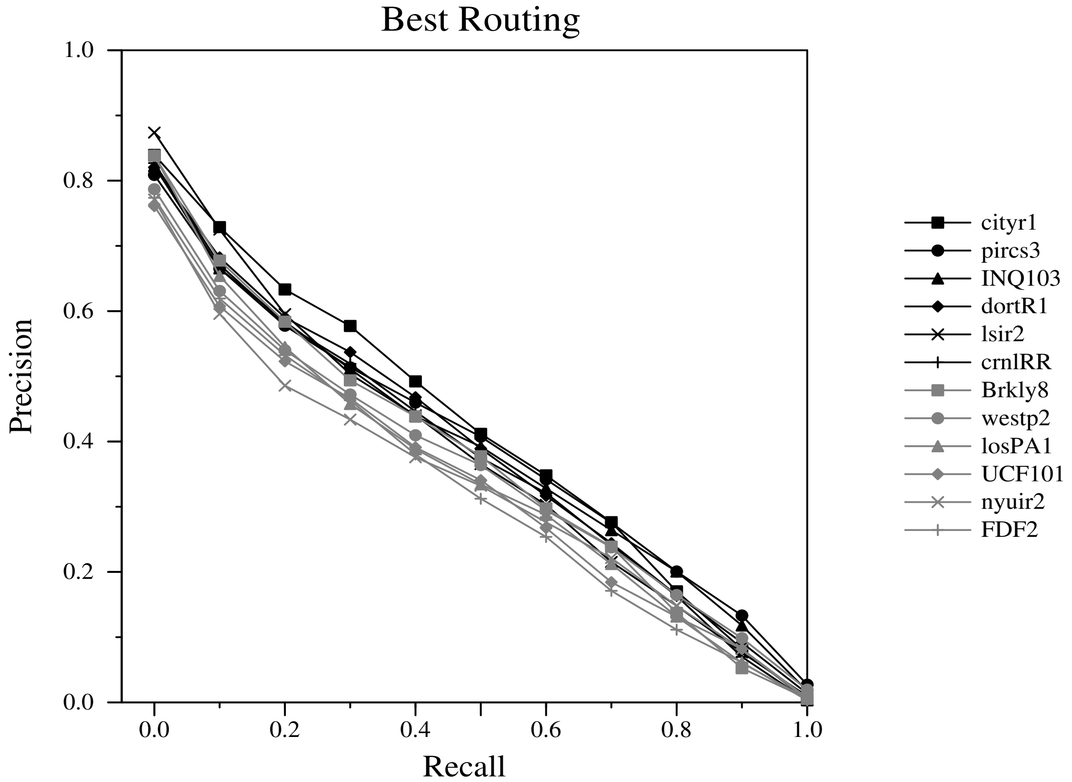

Figure 5. Best Routing Results.

westp2 -- West Publishing Company (see paper "TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System" by Paul Thompson, Howard Turtle, Bokyung Yang and James Flood) used their commercial product (WIN), but expanded the topics using up to 50 terms from specially selected parts of relevant documents.

losPA1 -- Logicon, Inc. (see paper "Research in Automatic Profile Creation and Relevance Ranking with LMDS" by Julian A. Yochum) constructed profiles based on the top 10 selected terms from the relevant documents, with term selection based on binomial probability distributions. The profile was used to select all documents containing any of those terms and the documents were then ranked using a weighting formula.

UCF101 -- University of Central Florida (see paper "Using Database Schemas to Detect Relevant Information" by James Driscoll, Gary Theis and Gene Billings) manually constructed entity-relationship schemas for each topic and also manually created synonym lists for each labelled component in the ER schema. These schemas and lists were then used to select and rank documents.

nyuir2 -- New York University (see paper "Natural Language Information Retrieval: TREC-3 Report" by Tomek Strzalkowski, Jose Carballo and Mihnea Marinescu) used NLP techniques to discover syntactic phrases in the documents. Both single terms and phrases were indexed and specially weighted. The *nyuir2* run used topic expansion based on the relevant documents.

FDF2 -- Paracel, Inc. (see paper "The FDF Query Generation Workbench" by K.I. Yu, P. Scheibe and F. Nordby) used a series of tools to generate profiles. These tools used statistical methods to create several alternative queries, and automatically evaluated the queries against the training data to select the best query for each topic.

The recall/precision curves shown in Figure 5 are very close in performance for the routing, with the Scheffé tests done by Jean Tague-Sutcliffe showing that there is no significant differences between the top 22 runs. It is, however, useful to look at the results on a per topic basis to find trends in performance across techniques.

The main issue for the TREC-3 routing runs is how to best select terms for topic expansion. Note that for the adhoc task the main issue was how to expand a topic beyond its original terms, hopefully with as little loss in precision as possible. For the routing task, however, the pool of terms for expansion is easily determined (i.e., the terms in the relevant documents), and the problem is how to select terms from this very large pool. Correspondingly, the major differences in results between the routing runs are not how many relevant documents were "missed" (as for the adhoc task), but how well the relevant documents were ranked.

An example of this is a comparison between the two City runs. The *cityr1* system used all relevant documents to select the top T terms, where T varied between 3 and 100 (average 47). Then they used the training material to optimize the queries, selecting only those terms that improved results. On average only about 17 terms were used in an optimized query. The unoptimized version of these queries was used at the *cityr2* run (not shown in Figure 5), which did not work as well. The difference in average precision between the two runs is only about 12%, but the optimized *cityr1* run had 14 superior topics (topics with a 20% or more improvement in average precision), all caused by better ranking (more relevant documents moved into the top 100 documents from the top 1000 documents). A similar comparison can be made between the *cityr1* run and the *pircs3* run. Even though there were more relevant documents found by the *pircs3* technique, the *cityr1* run had 15 superior topics (versus 7 superior for *pircs3*), all caused by better ranking.

The ability to assign better ranks to relevant documents is not strictly tied to being highly selective of terms. A comparison of the *cityr1*, *pircs3*, *INQ103* and *CrnlRR* runs shows that the INQUERY and PIRCS techniques both used an average of around 100 terms in their queries and retrieved the largest number of relevant documents in the top 1000 documents. The *cityr1* run, with only about 17 terms, missed a few relevant documents, but did a much better job of ranking the ones they found. However, even though the *CrnlRR* run used a massive expansion of greater than 300 terms, the *CrnlRR* runs were stronger in ranking than in finding relevant documents. A comparison of the *INQ103* run to that of Cornell shows that Cornell had 12 "inferior" topics, mostly due to missed relevant documents, and 9 superior topics, mostly due to better ranking. Clearly the appropriate number of terms to use in a routing query varies across retrieval techniques. This same result was seen in the adhoc task, where the appropriate number of expansion terms also varied across systems.

The top routing results tend to fall into three categories--those groups that used minimal effort in selecting terms (*CrnlRR*, *lsir2*), those groups that selected terms based on using only a portion of the relevant material (*pircs3* and *westp2*), and those groups that used all the material, but carefully selected terms (*cityr1*, *INQ103*, *brkly8* and *losPA1*).

Both the Cornell runs and the LSI runs were repeats of their TREC-2 techniques. The LSI runs tested using only the topic to create a query (no expansion) versus using all the relevant documents (no topic) to create a centroid for use as the query (the *lsir2* run). There is a 30% improvement using the relevant documents only. The Cornell runs used both the topic and a massive Rocchio relevance feedback expansion (300+ terms). Both groups used techniques based on a vector-space model (loosely based for the LSI technique), and this model appears to be able to effectively rank documents despite very massive queries. The strength of the Cornell ranking was mentioned before, but the LSI ranking is comparable or even better (18 superior topics for LSI, 9 for Cornell, all caused by better ranking).

Two groups (the PIRCS system and the WIN system from West) experimented with using only portions of the training data. This is mostly an efficiency issue, but also serves as a term selection method. The *pircs4* run (not shown in Figure 5) used only short documents, where short is defined as not more than 160 unique non-stop stems. This run did somewhat worse than the *pircs3* run, where a combination of these short documents and the top 2400 subdocuments were used. In both runs many fewer documents were used (12% and 35% of the relevant material respectively), yet the results were excellent. The West group tried multiple experiments using various segments of the relevant documents (best documents only, best 200 paragraphs, and best top paragraph). Up to 50 terms were added using a combination of the various approaches, with selection of approaches done on a per topic basis. This selective use of material caused some relevant documents to be missed. A comparison of the *westp2* run and the *INQ103* run shows that the 12 topics in which the *INQ103* run was superior were mostly caused by new relevant documents being found, whereas the 7 topics in which the *westp2* run was superior were all caused by better ranking.

Four groups (*cityr1*, *INQ103*, *brkly8*, and *LosPA1*) used all the relevant documents, but made careful selection of the terms to use. The City results have already been discussed. The *INQ103* run used an adaptation of the Rocchio algorithm with their inference engine technique. A statistical formula was used to select the top 32 terms to use for expansion for each topic, and then 30 additional terms were selected based on their proximity to those terms already selected. This technique retrieved a large number of the relevant documents into the top 1000 slots, but had more difficulties doing the ranking within that set. The *brkly8* run selected an average of over 1000 terms by using a chi-square test to indicate which stems were statistically associated with document relevance to a topic. These terms were weighted and used as the query. The *losPA1* run used a similar technique, calculating a binomial probability to select the top 1000 terms, selecting a pool of documents using an OR of the top 10 terms,

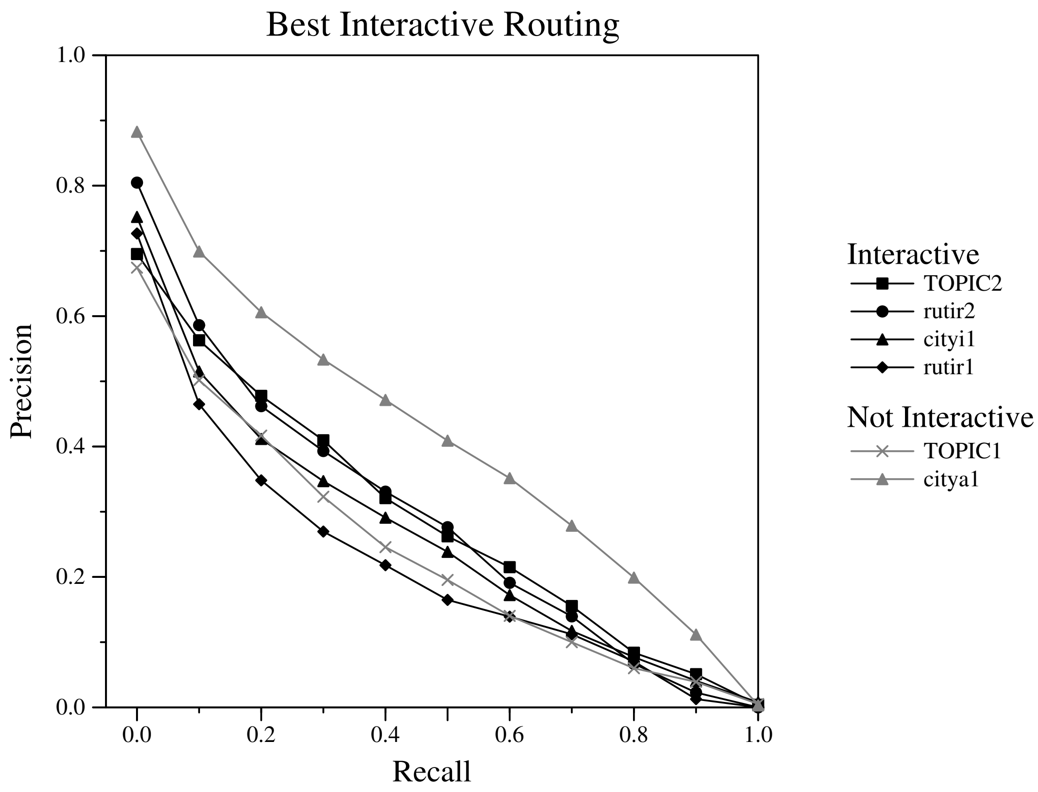

Figure 6. Interactive Results.

and then scoring the documents using a weighting algorithm based on occurrances of the 1000 terms in those documents. If results from these two systems are compared to the more traditional *INQ103* method, it seems that the strengths of these methods are in the ranking, with some problems in missing relevant documents.

As was the case in earlier TRECs, the manual construction of routing queries was not very competitive with automatic query construction. The manual *INQ104* run, consisting of a merge of the *INQ103* queries and a manually edited version of these queries was little different in results from the *INQ103* run. An exception to this was the reasonable results of the *UCF101* run. This run combined manually constructed detailed entity-relationship schema with manually constructed synonym lists. These were run against the documents, producing results that are comparable with the automatic results.

There is some improvement in overall routing results compared with those from TREC-2. This is mostly shown by the comparative position of the *CrnlRR* run, which was the "top-ranked" run in TREC-2, and now is more the "middle of the pack."

5.4 Other Experiments in TREC-3

In addition to the results aimed at producing high recall/precision performances, several groups did experiments using the TREC tasks to investigate other areas.

The largest area of experimentation was in interactive query construction, with four groups participating. One of the questions addressed by these groups was how well humans could perform the routing task, given a "rules-free" environment and access to the training material. The larger issue addressed by these experiments, however, was the entire interaction process in retrieval systems, since the "batch mode" evaluation of TREC does not reflect the way that most systems are used.

Figure 6 shows the three sets of results for the category A interactive runs, plus several baseline runs for comparison. A short summary of the systems follows, and readers are referred to the individual papers for more details.

TOPIC2 -- Verity, Inc. (see paper "Interactive Document Retrieval Using TOPIC (A report on the TREC-3 experiment)" by Richard Tong) used 12 Verity staff members ranging in search experience using TOPIC from novice to expert to build their queries. The initial queries were the manual-constructed queries used by Verity in TREC-2, and the results from these queries are shown in Figure 6 as *TOPIC1*. The searchers then improved the initial queries by periodically evaluating their "improved" queries against the training data. When sufficiently improved scores were achieved, the queries were declared final and used for TREC-3.

rutir1, *rutir2* -- Rutgers University (see paper "New Tools and Old Habits: The Interactive Searching Behavior of Expert Online Searches using INQUERY" by Jurgen Koenemann, Richard Quatrain, Colleen Cool and Nicholas Belkin) used the INQUERY system and had 10 experienced online searchers with no prior experience using that system build their queries. The entire query building process was restricted to 20 minutes per topic, and used the training data both for automatic relevance feedback (if desired) and for the searchers to check if a given retrieved document was relevant (as opposed to periodically evaluating their results). At some point during the 20 minute limit the queries were declared finished by the searchers and the results from these queries are shown in Figure 6 as *rutir1*. As a comparison, the experimenters also did the task themselves (*rutir2*).

cityi1 -- City University, London (see paper "Okapi at TREC-3" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used the OKAPI team as searchers. The initial query was manually generated using traditional operations. The retrieved documents (or a brief summary of them) were then displayed, and searchers checked the relevance judgments (generally viewing 10 or 12 relevant documents). Automatic relevance feedback was then applied and the searchers could choose to modify the resulting query or not (35 of the 50 topics were modified). Multiple iterations could be done before a decision was made on the final query.

Not shown in Figure 6 is a category B interactive result from the University of Toronto (see paper "Interactive Exploration as a Formal Text Retrieval Method: How Well can Interactivity Compensate for Unsophisticated Retrieval Algorithms" by Nipon Charoenkitkarn, Mark Chignell and Gene Golovchinsky). This group developed their TREC experiments from what was initially a browsing system. Boolean operators and promixity operators were used to construct the initial query. The queries were then "loosened" until around 1000 documents were retrieved. Then the results of these queries were run against the training data and reviewed, with changes possibly made to the query based on retrieval results.

As a group, the interactive results were considerably worse than the automatic routing results. This was somewhat unexpected since in all four cases the queries could be classified as the best manual queries possible. Although no definite reasons have been cited for this, the likely cause is the very strong performance of the automatic systems given the large amounts of training data.

A comparison of the City interactive run (*cityi1*) and the City automatic run (*citya1*) illustrates the problems. For BOTH runs, the query lengths were short, an average of around 17 terms. Only about 20% of these terms were in common, i.e., the searchers (*cityi1*) and the "computer" (*citya1*) picked different sets of terms. The difference in the results from these queries, however, is very large, as shown in Figure 6. The automatic run has a 63% improvement in average precision, and 33 topics with superior results (a 20% or more improvement in average precision) versus one topic with inferior results.

Regardless of the poorer performance, all four groups were able to draw interesting conclusions about their own interactive experiments. The Verity group found a 24% improvement in results (*TOPIC1* to *TOPIC2*) that can be obtained by humans using the training material over the (manually created) initial query. Other groups were able to gain insight into better tools needed by their system or insight into how online searchers handle the new techniques available. Of particular interest are the reports in these papers about the detailed human/computer interactions, as this provides insight on how systems might work in an operational setting.

A second area that drew more attention in TREC-3 was that of efficiency. Efficiency has always been an issue in TREC because sufficient efficiency (in both time and storage) is necessary to finish the tasks, and greater efficiency allows more experiments to be done within the same time period. Additionally the commercial systems in TREC must make any new algorithms fit into their already very efficient methodologies (the TRW/Paracel Fast Data Finder is a good example of these problems).

Many groups addressed efficiency issues in their TREC-3 papers, but the group from RMIT (see paper "Information Retrieval Systems for Large Document Collections" by Alistair Moffat and Justin Zobel) has specialized in efficiency issues in all the TRECs. In TREC-3 they investigated the issue of creating a centralized index in blocks for more efficient retrieval. They also tested text compression methods for dynamic document databases. Efficiency is likely to continue to be a major issue in TREC, possibly playing a larger part in the future.

A third area, that of properly handling heterogeneous collections such as the five main "subcollections" in TREC, was comprehensively addressed by the Siemens group (see paper "The Collection Fusion Problem" by Ellen Voorhees, Narendra Gupta and Ben Johnson-Laird). This group examined two different collection fusion techniques and was able to obtain results within 10% of the average precision of a run using a merged collection index. This type of investigation is important for real-world collections, and also to allow researchers to take advantage of possible variations in retrieval techniques for heterogeneous collections.

Several groups ran some experiments in thresholding as an alternative method of evaluating the routing task. For details on one of these experiments, see paper "TREC-2 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and D.D. Lewis.

The final set of experiments in TREC-3 involved starting work in a second language. Four groups worked with 25 topics in Spanish, using a document collection consisting of about 200 megabytes (58,000 records) of a Mexican newspaper from Monterey (*El Norte*). Since there was no training data for testing (similar to the startup problems for TREC-1), the groups used simple techniques.

CrnlVS, *CrnlES* -- Cornell University (see paper "Automatic Query Expansion Using SMART: TREC-3 by Chris Buckley, Gerard Salton, James Allan and Amit Singhal) used a baseline SMART run (*CrnlVS*) and a SMART run with massive topic expansion (*CrnlES*) similar to their English adhoc run. A simple stemmer and a stoplist of 342 terms were used.

SIN002, *SIN001* -- University of Massachusetts at Amherst (see paper "Document Retrieval and Routing Using the INQUERY System" by John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar) used the INQUERY system, with *SIN001* being a manually modified version of a basic automatic INQUERY run (*SIN002*) without topic expansion. A Spanish stemmer produced a 12% improvement in later experiments.

DCUSP1 -- Dublin City University (see paper "Indexing Structures Derived from Syntax in TREC-3: System Description" by Alan Smeaton, Ruairi O'Donnell and Fergus Kelledy) used a trigram retrieval model, with weighting of the trigrams from traditional frequency weighting. A Spanish stemmer based on the Porter algorithms was also used.

erims1 -- Environmental Research Institute of Michigan (see paper "Using an N-Gram-Based Document Representation with a Vector Processing Retrieval Model" by William Cavnar) used a quad-gram retrieval model, also with weighting using some of the traditional weighting mechanisms.

The major result from this very preliminary experiment in a second language was the ease of porting the retrieval techniques across languages. Cornell reported that only 5 to 6 hours of system changes were necessary (beyond creation of any stemmers or stopword lists).

6. Summary

The main conclusions that can be drawn from TREC-3 are as follows:

- Automatic construction of routers or filters from training data is very effective, much more effective than manual construction of these types of queries. This holds even if the manual construction is based on unrestricted use of the training data.
- Expansion of the shorter TREC-3 topics was highly successful, using either automatic topic expansion, manual topic expansion, or manually modified versions of automatically expanded topics. Many different techniques were effective, with research just beginning in this new area.
- The use of passage retrieval, subdocuments, and local weighting brings consistent performance improvements, especially in the adhoc task. Experiments this year show continued improvement coming from various methods of using these techniques to improve ranking.
- Preliminary results suggest that the extension of basic English retrieval techniques into another language (in particular Spanish) does not appear difficult. TREC-3 represents the first large-scale test of this portability issue.

Do these conclusions hold in the real world of text retrieval? Certainly the use of automatic construction of routers will work in any environment having reasonable amounts of training material. Of greater question is the transferability of the adhoc results. Two particular issues need to be addressed here. First, even though the topics in TREC-3 are shorter, they are still considerably longer than most queries used in operational settings. A couple of sentences is likely to be the maximum a user is willing to type into a computer, and it is unclear if the TREC topic expansion methods would work on these shorter input strings. Shorter topics may also need different techniques of passage retrieval and local weighting. TREC-4 will address this issue by using appropriately shorter topics.

The second mismatch of the TREC-3 results to the real-world is the emphasis on high recall in TREC. Requesting 1000 ranked documents and calculating the results on these goes well beyond average user needs. Karen Sparck Jones addresses this issue by looking at retrieval performance based only on the top 30 documents retrieved [Sparck Jones 1995, updated for TREC-3 in Appendix C to the proceedings]. An improvement of 20% in precision at this cutoff means that six additional relevant documents will be returned to the user, and this is likely to be noticeable by many users. Many of the techniques used in

TREC produced this difference; additionally some of the tools being investigated in TREC, such as the topic expansion tools, will make query modification much easier for the average user.

There will be a fourth TREC conference in 1995, and most of the systems that participated in TREC-3 will be back, along with additional groups. The routing and adhoc tasks will be done again, with different data and even shorter adhoc topics. In addition special new tasks (call "tracks") will be created to provide a focus to those areas of TREC that have been attracting more experimental interest. Six tracks will be tried.

- Interactive -- investigating searching as an interactive task by examining the process as well as the outcome.
- Multilingual -- working with non-English test collections (250 megabytes of Spanish and 25 topics, plus possibly Chinese and/or Japanese collections).
- NLP -- more focussed investigation of NLP in an IR environment, emphasizing the discovery and use of phrases for TREC-4.
- Multiple database merging -- investigation of techniques for merging results from the various TREC subcollections.
- Data corruption -- examining the effects of corrupted data (such as would come from an OCR environment) by using corrupted versions of the TREC data.
- Filtering -- evaluating routing systems on the basis of retrieving an unranked set of documents optimizing a specific effectiveness measure.

Acknowledgments

The author would like to gratefully acknowledge the continued support of the Software and Intelligent Systems Technology Office of the Advanced Research Projects Agency for the TREC conferences. Special thanks also go to the TREC program committee and the staff at NIST.

7. References

Harman D. (Ed.). (1993). *The First Text REtrieval Conference (TREC-1).* National Institute of Standards and Technology Special Publication 500-207, Gaithersburg, Md. 20899.

Harman D. (Ed.). (1994a). *The Second Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology Special Publication 500-215, Gaithersburg, Md. 20899. Also a special issue of *Information Processing and Management*, 31(3) in press.

Harman D. (1994b). Data Preparation. In: Merchant R. (Ed.).*The Proceedings of the TIPSTER Text Program - Phase I.* San Mateo, California: Morgan Kaufmann Publishing Co., 1994.

Sparck Jones K. and Van Rijsbergen C. (1975). *Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection*, British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.

Sparck Jones K. (1995). Reflections on TREC. *Information Processing and Management*, 31(3), in press.