Overview of the First Text REtrieval Conference (TREC-1)

Donna Harman
National Institute of Standards and Technology
Gaithersburg, Md. 20899

1. Introduction

There is a long history of experimentation in information retrieval. Research started with experiments in indexing languages, such as the Cranfield I tests (Cleverdon 1962), and has continued with over 30 years of experimentation with the retrieval engines themselves. The Cranfield II studies (Cleverdon et al. 1966) showed that automatic indexing was comparable to manual indexing, and this and the availability of computers created a major interest in the automatic indexing and searching of texts. The Cranfield experiments also emphasized the importance of creating test collections and using these for comparative evaluation. The Cranfield collection, created in the late 1960's, contained 1400 documents and 225 queries, and has been heavily used by researchers since then. Subsequently other collections have been built, such as the CACM collection (Fox 1983), and the NPL collection (Sparck Jones & Webster 1979).

In the 30 or so years of experimentation there have been two missing elements. First, although some research groups have used the same collections, there has been no concerted effort by groups to work with the same data, use the same evaluation techniques, and generally compare results across systems. The importance of this is not to show any system to be superior, but to allow comparison across a very wide variety of techniques, much wider than only one research group would tackle. Karen Sparck Jones in 1981 commented that:

> Yet the most striking feature of the test history of the past two decades is its lack of consolidation. It is true that some very broad generalizations have been endorsed by successive tests: for example...but there has been a real failure at the detailed level to build one test on another. As a result there are no explanations for these generalizations, and hence no means of knowing whether improved systems could be designed (p. 245).

This consolidation is more likely if groups can compare results across the same data, using the same evaluation method, and then meet to discuss openly how methods differ.

The second missing element, which has become critical in the last 10 years, is the lack of a realistically-sized test collection. Evaluation using the small collections currently available may not reflect performance of systems in large full-text searching, and certainly does not demonstrate any proven abilities of these systems to operate in real-world information retrieval environments. This is a major barrier to the transfer of these laboratory systems into the commercial world. Additionally some techniques such as the use of phrases and the construction of automatic thesauri seem intuitively workable, but have repeatedly failed to show improvement in performance using the small collections. Larger collections might demonstrate the effectiveness of these procedures.

The overall goal of the Text REtrieval Conference (TREC) was to address these two missing elements. It is hoped that by providing a very large test collection, and encouraging interaction with other groups in a friendly evaluation forum, a new thrust in information retrieval will occur. There is also an increased interest in this field within the DARPA community, and TREC is designed to be a showcase of the state-of-the-art in retrieval research. NIST's goal as co-sponsor of TREC is to encourage communication and technology transfer among academia, industry, and government.

2. The Task

2.1 Introduction

TREC is designed to encourage research in information retrieval using large data collections. Two types of retrieval are being examined -- retrieval using an "adhoc" query such as a researcher might use in a library environment, and retrieval using a "routing" query such as a profile to filter some incoming document stream. The TREC task is not tied to any given application, and is not concerned with interfaces or optimized response time for searching. However it is helpful to have some potential user in mind when designing or testing a retrieval system. The model for a user in TREC is a dedicated searcher, not a novice searcher, and the model for the application is one needing monitoring of data streams for information on specific topics (routing), and the ability to do adhoc searches on archived data for new topics. It should be assumed that the users need the ability to do both high precision and high recall searches, and are willing to look at many documents and repeatedly modify queries in order to get high recall. Obviously they would like a system that makes this as easy as possible, but this ease should be reflected in TREC as added intelligence in the system rather than as special interfaces.

Since TREC has been designed to evaluate system performance both in a routing (filtering or profiling) mode, and in an adhoc mode, both functions need to be tested. The test design was based on traditional information retrieval models, and evaluation used traditional recall and precision measures. The following diagram of the test design shows the various components of TREC (fig. 1).

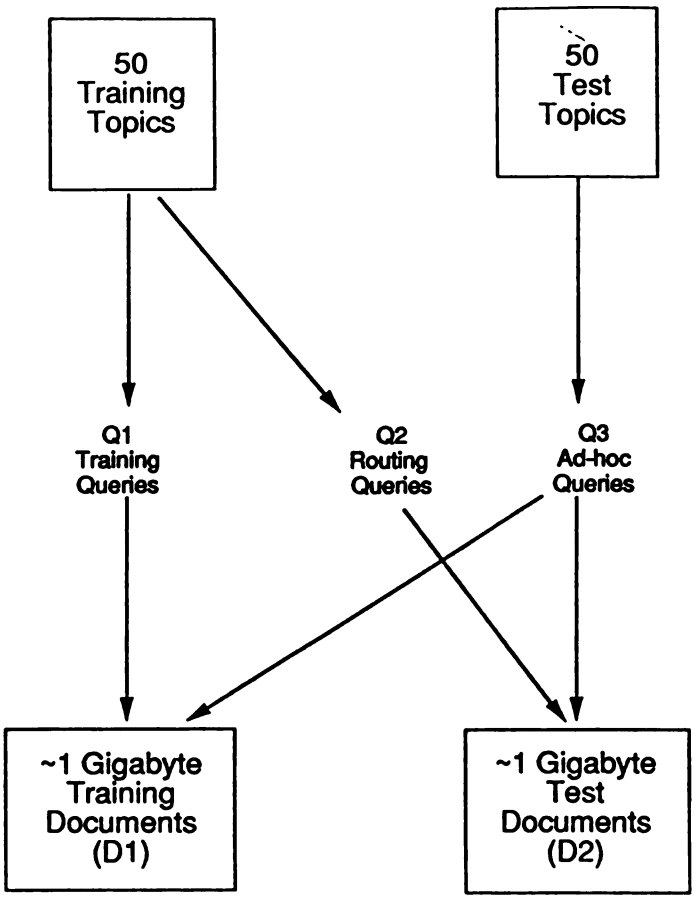

Figure 1. The TREC Task.

This diagram reflects the four data sets (2 sets of topics and 2 sets of documents) that were provided to participants. These data sets (along with a set of sample relevance judgments for the 50 training topics) were used to

construct three sets of queries. Q1 is the set of queries (probably multiple sets) created to help in adjusting a system to this task, to create better weighting algorithms, and in general to train the system for testing. The results of this research were used to create Q2, the routing queries to be used against the test documents. Q3 is the set of queries created from the test topics as adhoc queries for searching against the combined documents (both training documents and test documents). The results from searches using Q2 and Q3 were the official test results. The documents were full-length text from various sources such as newspapers, newswires, magazines and journals (see sect. 3.2 for more details).

2.2 Specific Task Guidelines

The various TREC participants used a wide variety of indexing/knowledge base building techniques, and a wide variety of approaches to generate search queries. Therefore it was important to establish clear guidelines for the TREC task and to develop some methods of standardized reporting to allow comparison. The 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 were constructed to reflect an actual operational environment, and to allow as fair as possible a separation among the diverse query construction approaches.

There were guidelines for constructing and manipulating the system data structures. These structures were defined to consist of the original documents, any new structures built automatically from the documents (such as inverted files, thesauri, conceptual networks, etc.) and any new structures built manually from the documents (such as thesauri, synonym lists, knowledge bases, rules, etc.). The following guidelines were provided to the participants.

1. System data structures can be built using the initial training set (documents D1, training topics, and relevance judgments). They may be modified based on the test documents D2, but not based on the test topics. In particular, the processing of one test topic should not affect the processing of another test topic. For example, it would not be allowed to update a system knowledge base based on the analysis of one test topic in such a way that the interpretation of subsequent test topics was changed in any fashion.
2. There are several parts of the Wall Street Journal and the Ziff material (see sect. 3.2) that contain manually assigned controlled or uncontrolled index terms. These fields are delimited by SGML tags, as specified in the documentation files included with the data. Other parts of the TREC data contain no manual indexing. Since the primary focus of TREC is on retrieval and routing of naturally occurring text, these manually indexed terms should not be indiscriminately used as if they are a normal part of the text. If your group decides to use these terms, they should be part of a specific experiment that utilizes manual indexing terms, and their use should be declared.
3. Special care should be used in handling the routing topics. In a true routing situation, a single document would be indexed and "passed" against the routing topics. Since most of you will be indexing the test documents as a complete set, routing should be simulated by not using any test document information (such as IDF based on the test collection, total frequency based on the test collection, etc.) in the searching. It is perfectly permissible to use training-set collection information however. If your system bases system data structures on the entire test data and is unable to operate in a proper routing mode, then you should either have a different method for handling routing, or only submit results for the adhoc part of TREC.

Additionally there were guidelines for constructing the queries from the provided topics (see sect. 3.3 for more on the topics). These guidelines were considered of great importance for fair system comparison and were therefore carefully constructed. Three generic categories were defined, based on the amount and kind of manual intervention used.

1. Method 1 -- completely automatic initial query construction.

   adhoc queries -- The system will automatically extract information from the topic (the topic fields used should be identified) to construct the query. The query will then be submitted to the system (with no manual modifications) and the results from the system will be the results submitted to NIST. There should be no manual intervention that would affect the results.

   routing queries -- The queries should be constructed automatically using the training topics, the training relevance judgments and the training documents. The queries should then be submitted to NIST before the

test documents are released and should not be modified after that point. The unmodified queries should be run against the test documents and the results submitted to NIST.

2. Method 2 -- manual initial query construction.

   adhoc queries -- The query is constructed in some manner from the topic, either manually or using machine assistance. The methods used should be identified, along with the human expertise (both domain expertise and computer expertise) needed to construct a query. Once the query has been constructed, it will be submitted to the system (with no manual intervention), and the results from the system will be the results submitted to NIST. There should be no manual intervention after initial query construction that would affect the results. (Manual intervention is covered by Method 3.)

   routing queries -- The queries should be constructed in the same manner as the adhoc queries for Method 2, but using the training topics, relevance judgments, and training documents. They should then be submitted to NIST before the test documents are released and should not be modified after that point. The unmodified queries should be run against the test documents and the results submitted to NIST.

3. Method 3 -- automatic or manual query construction with feedback.

   adhoc queries -- The initial query can be constructed using either Method 1 or Method 2. The particular technique used should be described. The query is submitted to the system, and a subset of the retrieved documents is used for manual feedback, i.e., a human makes judgments about the relevance of the documents in this subset. These judgments may be communicated to the system, which may automatically modify the query, or the human may simply choose to modify the query himself. In either case, the expertise of the person or persons examining the documents should be described, both their domain expertise and their experience in online searching, and the manner of system feedback (i.e., automatic system modification of query or human modification) should be also described. At some point, feedback should end, and the query should be accepted as final.

   Three sets of results should be sent to NIST for each topic. The first set should be the results without feedback, i.e., the top 200 documents retrieved from an initial query produced without feedback, whether produced manually or automatically. This set should be exactly the same as the results from Method 1 or Method 2, but should be submitted again as one part of Method 3. The second set should be the results after only one iteration of feedback, with the top X documents used in the first iteration of feedback frozen. For example, if you "used" the top 20 documents for feedback, then the second set of results should have these documents as the top 20 documents, followed by the top 200 documents retrieved based on feedback. The term "used" means all documents for which some information has been seen by the judger, and are deemed by the system to have been seen. These two sets of results will be used by NIST to calculate a residual evaluation measure. The third set of results should be a record of your feedback, i.e., a list of documents in the exact order they were seen and judged, with an indication of iteration boundaries. For example, if you ran six iterations of feedback, with 10 documents looked at for each iteration, the record would be a list of the 60 documents seen by the "user", marked at 10, 20, 30, etc. You should also indicate what information the user communicated to your system about each document (relevant/not relevant, too general/too specific/on target, etc.). We will be specifying a format for these record files later. These files will be used to calculate measures based on the total number of relevant documents retrieved both across iterations and across a given document level.

   routing queries -- Method 3 cannot be used for routing queries because routing systems have typically not supported feedback.

In general these guidelines served well, although there was some misunderstanding about what constituted feedback. The guidelines will be clarified for TREC-2.

2.3 The Participants

There were 25 participating systems in TREC-1, using a wide range of retrieval techniques. The participants were able to choose from three levels of participation: Category A, full participation, Category B, full participation using a reduced dataset (25 topics and 1/4 of the full document set), and Category C for evaluation only (to allow commercial systems to protect proprietary algorithms). The program committee selected only 20 category A and B groups to present talks because of limited conference time, and requested that the rest of the groups present posters. All groups were asked to submit papers for the proceedings.

Each 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 (Methods 1, 2, or 3), 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 200 documents retrieved for each topic for evaluation.

3. The Test Collection

3.1 Introduction

Critical to the success of TREC was the creation of the test collection. Like most traditional retrieval collections, there are three distinct parts to this collection. The first is the documents themselves -- the training set (D1) and the test set (D2). Both were distributed as CD-ROMs with about 1 gigabyte of data each, compressed to fit. The training topics, the test topics and the relevance judgments were supplied by email. TREC-1 used the same test collection (documents and topics) used in the DARPA TIPSTER project. (The DARPA TIPSTER project involves the same tasks as TREC, but with four contractors doing more intense research than is being expected from TREC participants (Harman 1993)). However a major increase in the number of relevance judgments for this collection became available from the TREC-1 evaluation.

The components of the test collection -- the documents, the topics, and the relevance judgments, are discussed in the rest of this section.

3.2 The Documents

The documents came from the following sources.

- Disk 1
  - WSJ -- Wall Street Journal (1986, 1987, 1988, 1989)
  - AP -- AP Newswire (1989)
  - ZIFF -- Information from Computer Select disks (Ziff-Davis Publishing)
  - FR -- Federal Register (1989)
  - DOE -- Short abstracts from the Department of Energy
- Disk 2
  - WSJ -- Wall Street Journal (1990, 1991, 1992)
  - AP -- AP Newswire (1988)
  - ZIFF -- Information from Computer Select disks (Ziff-Davis Publishing)
  - FR -- Federal Register (1988)

The particular sources were selected because they reflected the different types of documents used in the imagined TREC application. Specifically they had a varied length, a varied writing style, a varied level of editing and a varied vocabulary. All participants were required to sign a detailed user agreement for the data in order to protect the copyrighted source material.

The documents were uniformly formatted into an SGML-like structure, as can be seen in the following example.

<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 implications for computer and communications equipment markets.

AT&T said it is the first national long-distance carrier to announce prices for specific services under a world-wide standardization plan to upgrade phone networks. By announcing commercial services under the plan, which the industry calls the Integrated Services Digital Network, AT&T will influence evolving communications standards to its advantage, consultants said, just as International Business Machines Corp. has created de facto computer standards favoring its products.

.
.
.
</TEXT>
</DOC>

All documents had beginning and end markers, and a unique DOCNO id field. Additionally other fields taken from the initial data appeared, but these varied widely across the different sources. The documents also had different amounts of errors, which were not checked or corrected. Not only would this have been an impossible task, but the errors in the data provided a better simulation of the real-world task. Errors in missing document separators or bad document numbers were screened out, although a few were missed and later reported by participants.

Table 1 shows some basic document collection statistics.

TABLE 1. DOCUMENT STATISTICS

Subset of collection	WSJ	AP	ZIFF	FR	DOE
Size of collection (megabytes)					
(disk 1)	295	266	251	258	190
(disk 2)	255	248	188	211	
Number of records					
(disk 1)	98,736	84,930	75,180	26,207	226,087
(disk 2)	74,520	79,923	56,920	20,108	
Median number of terms per record					
(disk 1)	182	353	181	313	82
(disk 2)	218	346	167	315	
Average number of terms per record					
(disk 1)	329	375	412	1017	89
(disk 2)	377	370	394	1073	

Note that although the collection sizes are roughly equivalent in megabytes, there is a range of document lengths 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 and ZIFF documents have a wider range of lengths. The documents from the Federal Register (FR) have a very wide range of lengths.

The distribution of terms in these subsets show interesting variations. Table 2 shows some term distribution statistics found using a small stopword list of 25 terms and no stemming. For example the AP has more unique terms than the others, probably reflecting both more proper names and more spelling errors. The DOE collection, while very small, is highly technical and has many domains, resulting in many specific technical terms.

TABLE 2. DICTIONARY STATISTICS

Subset of collection	WSJ	AP	ZIFF	FR	DOE
Total number of unique terms					
(disk 1)	156,298	197,608	173,501	126,258	186,225
(disk 2)	153,725	186,500	147,405	116,586	
Occurring once					
(disk 1)	64,656	89,627	85,992	58,677	95,782
(disk 2)	64,844	83,019	72,053	54,823	
Occurring more > 1					
(disk 1)	91,642	107,981	87,509	67,581	90,443
(disk 2)	88,881	103,481	75,352	61,763	
Average number of occurrences > 1					
(disk 1)	199	174	165	106	159
(disk 2)	178	169	139	91	

How does this document set compare with the older collections? Table 3 shows a comparison of these collections with the Cranfield 1400 collection mentioned earlier. Not only has the size of the document collection increased by a factor of about 200, but the average length of the documents has at least doubled, and in some cases (FR), increased by a factor of 10. Also, the dictionary sizes have increased by a factor of 20.

TABLE 3. COMPARISON TO OLDER COLLECTIONS

	Subset of collection				
Size of collection (megabytes)	295	266	251	258	1.5
Number of records	98,736	84,930	75,180	26,207	1400
Median number of terms per record	182	353	181	313	79
Average number of terms per record	329	375	412	1017	88
Total number of unique terms	156,298	197,608	173,501	126,258	8226

What does this mean to the TREC task? First, a major portion of the effort for TREC-1 was spent in the system engineering necessary to handle the huge number of documents. This means that little time was left for system tuning or experimental runs, and therefore the TREC-1 results can best be viewed as a baseline for later research. The longer documents also required major adjustments to the algorithms themselves (or loss of performance). This is particularly true for the very long documents in FR. Since a relevant document might contain only one or two relevant sentences, many algorithms needed adjustment from working with the abstract length documents found in the old collections. Additionally many documents were composite stories, with different topics, and this caused problems for most algorithms.

3.3 The Topics

In designing the TREC (and TIPSTER) tasks, 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 topics were designed to mimic a real user's need, and were written by people who are actual users of a retrieval system. Although the subject domain of the topics was diverse, some consideration was given to the documents to be searched. The topics were constructed by doing trial retrievals against a sample of the document set, and then those topics that had roughly 25 to 100 hits in that sample were used. This created a range of broader and narrower topics.

The following is one of the topics used in TREC.

<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing

<desc> Description:
Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.

<narr> Narrative:
A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one or more features of the company's product.

<con> Concept(s):
1. natural language processing
2. translation, language, dictionary, font
3. software applications

<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>

Each topic was formatted in the same standard method to allow easier automatic construction of queries. Besides a beginning and an end marker, each topic had a number, a short title, and a one-sentence description. There was a narrative section which was aimed at providing a complete description of document relevance for the assessors. Each topic also had a concepts section with a list of assorted concepts related to the topic. This section was designed to provide a mini-knowledge base about a topic such as a real searcher might possess. Additionally each topic could have a definitions section and/or a factors section. The definition section had one or two of the definitions critical to a human understanding of the topic. The factors section was included to allow easier automatic query building by listing specific items from the narrative that constrain the documents that are relevant. Two particular factors were used in the TREC-1 topics: a time factor (current, before a given date, etc.) and a nationality factor (either involving only certain countries or excluding certain countries).

While the TREC topics did not present a problem in scaling, the challenge of either automatically constructing a query, or manually constructing a query with little foreknowledge of its searching capability, was a major challenge for TREC participants. In addition to filtering the relatively large amount of information provided in the topics into queries, the sometimes narrow definition of relevance as stated in the narrative was difficult for

most systems to handle. The two narratives shown below illustrate this point.

<num> Number: 051

A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.

<num> Number: 058

A relevant document will either report an impending rail strike, describing the conditions which may lead to a strike, or will provide an update on an ongoing strike. To be relevant, the document will identify the location of the strike or potential strike. For an impending strike, the document will report the status of negotiations, contract talks, etc. to enable an assessment of the probability of a strike. For an ongoing strike, the document will report the length of the strike to the current date and the status of negotiations or mediation.

In a preliminary analysis, the narratives and the factors played a strange and unpredictable role in the results for TREC-1. Systems did as well on topics with very restrictive narratives, such as that of topic 58, as on topics with non-restrictive narratives, such as topic 51. The subject and terms in the entire topic were more important in determining success than the restrictiveness of the narrative. The factors also did not play a major role in system performance. This could change in TREC-2 when groups have more time to adjust their systems to the TREC task.

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. For the TREC task, three possible methods for finding the relevant documents could have been used. In the first method, full relevance judgments could have been made on all 742,611 documents, for each topic, resulting in over 74 million judgments. This was clearly impossible. As a second approach, a random sample of the documents could have been taken, with relevance judgments done on that sample only. The problem with this approach is that a random sample that is large enough to find on the order of 200 relevant documents per topic is a very large random sample, and is likely to result in insufficient relevance judgments. The third method, the one used in TREC, was to make relevance judgments on the sample of documents selected by the various participating systems. This method is known as the pooling method, and has been used successfully in creating other collections. It was the recommended method in 1975 proposal to the British Library to build a very large test collection (Sparck Jones & van Rijsbergen).

To construct the pool, the following was done.

1. Divide each set of results into results for a given topic
2. For each topic within a set of results, select the top 200 ranked documents for input to the pool
3. For each topic, merge results from all systems
4. For each topic, sort results based on document numbers
5. For each topic, remove duplicate documents

Pooling proved to be an effective method. There was little overlap among the 25 systems in their retrieved documents. Table 4 shows the overlap statistics. The first overlap statistics are for the adhoc topics (test topics against both training documents D1 and test documents D2), and the second statistics are for the routing topics (training topics against test documents D2 only).

TABLE 4. OVERLAP OF SUBMITTED RESULTS				
	Top 200 Possible	Top 200 Actual	Top 100 Possible	Top 100 Actual
Average Number of Unique Documents Per Topic (Adhoc, 33 runs, 16 groups)	6600	2398.4	3300	1278.86
Average Number of Unique Documents Per Topic (Routing, 22 runs, 16 groups)	4400	1932.42	2200	1066.86

For example, out of a maximum of 6600 unique documents (33 groups times 200 documents), over one-third were actually unique. The top 100 documents retrieved contained about the same percentage of unique documents. This means that the different systems were finding different documents as likely relevant documents for a topic. Whereas this might be expected (and indeed has been shown to occur, Katzer et. al. 1982) from widely differing systems, these overlaps were often between two runs for a given system, or between two systems run on the same basic retrieval engine. One reason for the lack of overlap is the very large number of documents that contain many of the same keywords as the relevant documents, but probably a larger reason is the very different sets of keywords in the constructed queries (this needs further analysis). This lack of overlap should improve the coverage of the relevance set, and verifies the use of the pooling methodology to produce the sample.

The merged list of results was then shown to the human assessors. Only the top 100 documents were judged, resulting in an average of 1462.24 documents judged for each topic, and ranging from a high of 2893 for topic 74 to a low of 611 for topic 46. Each topic was judged by a single assessor to insure the best consistency of judgment and varying numbers of documents were judged relevant to the topics. Figure 3 shows the number of documents judged relevant for each of the 100 topics. The topics are sorted by the number of relevant documents to better show their range and median.

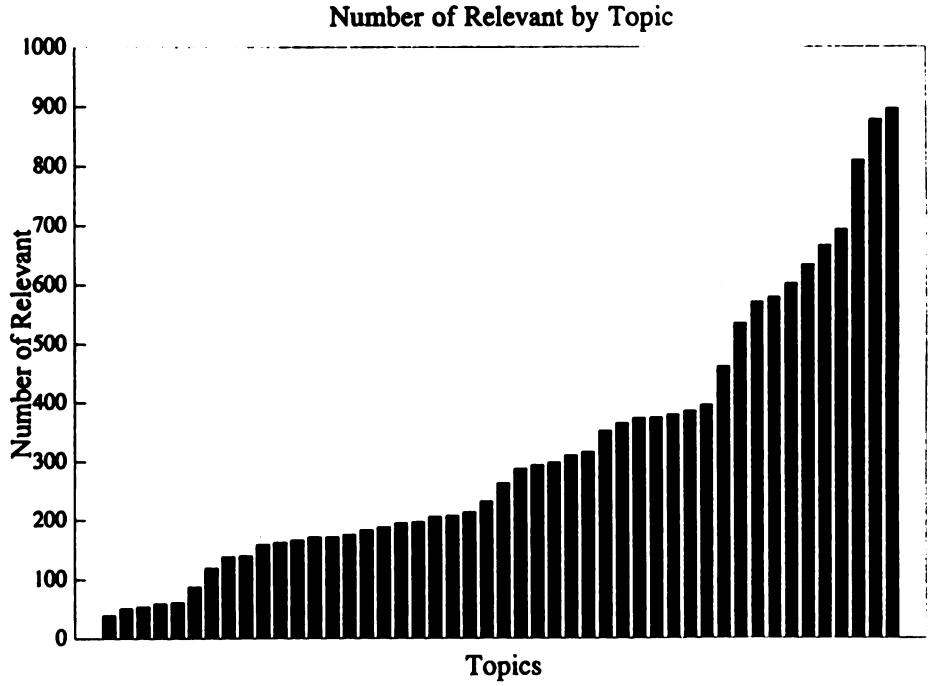

Figure 3. Number of Relevant Documents on a Per Topic Basis.

4. Evaluation

4.1 Existing Evaluation Methodology

An important element of TREC was to provide a common evaluation forum. Standard recall/precision figures were calculated for each system and the tables and graphs for the results are presented in Appendix A. Figure 4 shows a typical recall/precision curve for illustration purposes. The x axis plots the recall values at fixed levels of recall, where

$$Recall = \frac{number\ of\ relevant\ items\ retrieved}{total\ number\ of\ relevant\ items\ in\ collection}$$

The y axis plots the average precision values at those given recall values, where precision is calculated by

$$Precision = \frac{number\ of\ relevant\ items\ retrieved}{total\ number\ of\ items\ retrieved}$$

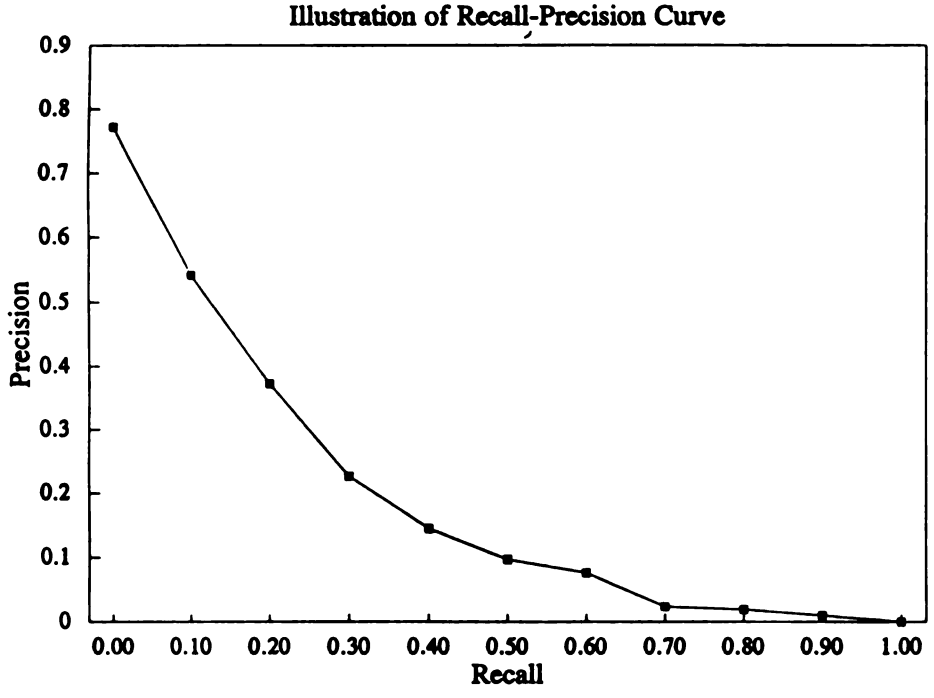

Figure 4. A Sample Recall/Precision Curve.

There is a standard table and graph in Appendix A for each run for each system, with the runs identified by their unique tags. A map for matching the tags to the systems is also provided. Note that the tables for the TIPSTER panel are in Appendix B as the results are not directly comparable to the TREC results. The tables show some total statistics for each run, plus both the recall-level and document-level recall/precision averages.

A second type of information about each system is shown in Appendix C. These standardized forms describe system features and system timing, and allow some primitive comparison of the amount of effort needed to produce the results.

4.2 Problems with Evaluation

Since this was the first time that such a large collection of text has been used in evaluation, there were some problems using the existing methods of evaluation. First, groups were asked to send in only the top 200 documents retrieved by their systems. This artificial document cutoff is relatively low and systems did not retrieve all the relevant documents for most topics within the cutoff. All documents retrieved beyond the 200 were considered nonrelevant by default and therefore the recall/precision curves become inaccurate after about 40% recall on average. Table 5 shows a comparison of one system using no threshold (so relevant documents found beyond the 200 limit are marked as relevant) versus using the 200 document threshold.

TABLE 5. COMPARISON OF TABLES FROM TIPSTER

Full Ranking		Top 200 Ranking	
Recall	Precision	Precision	
0.0	0.821	0.8208	
0.1	0.672	0.6710	
0.2	0.581	0.5759	
0.3	0.528	0.5030	
0.4	0.472	0.3819	
0.5	0.424	0.2999	
0.6	0.368	0.1773	
0.7	0.315	0.1075	
0.8	0.244	0.0487	
0.9	0.154	0.0117	
1.0	0.039	0.0000	
11 pt. average	0.421	0.3271	
Recall	Precision	Recall	Precision
0.25	0.559	0.20	0.5759
0.50	0.424	0.50	0.2999
0.75	0.280	0.80	0.0487
3 pt. average	0.421		0.3082

It can be seen from these tables that not only are the recall-level statistics beyond about 40% recall inaccurate, but both the 11 pt. and the 3 pt. averages based on this table are also inaccurate. Since all systems were compared using the same measures, this problem is not serious in terms of comparing methods within TREC-1. However, it could be improved by lowering the threshold, and TREC-2 will be run such that at least the top 500 documents are used for evaluation.

A related problem occurred because some systems in TREC-1 worked on a variable thresholding system, with that threshold set for each topic. Documents not matching sufficient system criteria were rejected, even if fewer than 200 were returned. Sometimes as few as 10 documents were sent as results, and the evaluation method again assumed all documents beyond the 10 were not relevant. This hurt performance for these systems badly in some cases and the individual system papers discuss this. The plans for TREC-2 are to include some additional thresholding tests, so that these systems can evaluate how their thresholding performs and evaluate the standard ranking as done by other systems.

The third problem was more general in nature. The current recall/precision measures do not include any indication of the collection size. This means that the recall and precision of a system based on a 1400 document collection could be the same as that of a system based on a million document collection, but obviously the discrimation powers on a million document collection would be much greater. This may not have been a problem on the smaller collections, but the discrimation power of systems on TREC-sized collections is very important. Clearly some new evaluation measures are needed for this.

One new measure being tried in TREC-1 is the ROC (Relative Operating Characteristic) curves used in signal processing. These curves are similar to the recall/precision curves, but allow the total size of the collection to influence performance. The two variables being used here are the probability of detection or probability of a

"hit" versus the probability of false alarm, or the probability of a "false drop". The x axis plots the probability of false alarm, calculated as follows

$$\textit{Probability of false alarm} = \frac{\textit{number of nonrelevant items retrieved}}{\textit{total number of nonrelevant items in collection}}$$

The y axis plots the probability of detection, calculated as

$$\textit{Probability of detection} = \frac{\textit{number of relevant items retrieved}}{\textit{total number of relevant items in collection}}$$

Note that the probability of detection is the same as recall, and the probability of false alarm is the same as fallout, an older measure in information retrieval (Salton & McGill 1983). These measures are for a single topic, but averages can be computed similarly to the recall-level averages by using probability of detection at fixed false alarm rates. The tables in Appendix A show both this average ROC curve and the same curve plotted on probability scales (Swets 1969).

5. Preliminary Results

5.1 Introduction

The results of the TREC-1 conference should be viewed only as a preliminary baseline for what can be expected from systems working with large test collections. There are several reasons for this. First, the deadlines for results were very tight, and most groups had minimal time for experiments. As discussed earlier, the huge scale-up in the size of the document collection required major work from all groups in rebuilding their systems. Much of this work was simply a system engineering task: finding reasonable data structures to use, getting indexing routines to be efficient enough to finish indexing the data, finding enough storage to handle the large inverted files and other structures, etc.

The second reason these results are preliminary is that groups were working blindly as to what constitutes a relevant document. There were no reliable relevance judgments for training, and the use of the long topics was completely new. This means that results were heavily influenced by an almost random selection of what parts of the topic to use. Groups also had to make often primitive adjustments to basic algorithms in order to get results, with little evidence of how well these adjustments were working. The large scale of the whole evaluation precluded any tuning without some relevance judgments, and the relevance judgments that were provided were generally sparse and sometimes inaccurate. These problems particularly affected those systems that needed training for routing.

Many of the papers in the proceedings show some new results from work done in the short amount of time between the conference and the due date of the papers (less than 2 months). Some of the improvements are very significant, and the improvements seen in the TIPSTER results (where the results are a second-try at this task) are large. It can be expected that the results seen at the second TREC conference will be much better, and also more indicative of how well a method works.

Because these results are preliminary, they should be compared very carefully. Some very broad conclusions can be drawn, but no methods should be conclusively judged inferior or superior at this point.

5.2 Adhoc Results

The adhoc evaluation used new topics (51-100) against the two disks of documents (D1 + D2). There were 33 sets of results for adhoc evaluation in TREC, with 20 of them based on runs for the full data set. Of these, 13 used automatic construction of queries, 6 used manual construction, and 1 used feedback. Figure 5 shows the recall/precision curve for the three TREC-1 runs with the highest 11-point averages using automatic construction of queries. These curves were all based on the use of the Cornell SMART system, but with important variations. The "fuhrp1" results came from using the training data to find parameter weights (see Fuhr & Buckley paper), the "crnlp1" results came from doing local and global term weighting without training data (see Buckley, Salton & Allan paper), and the "siems1" results came from using term expansion with terms from "Wordnet" (see Voorhees & Hou paper).

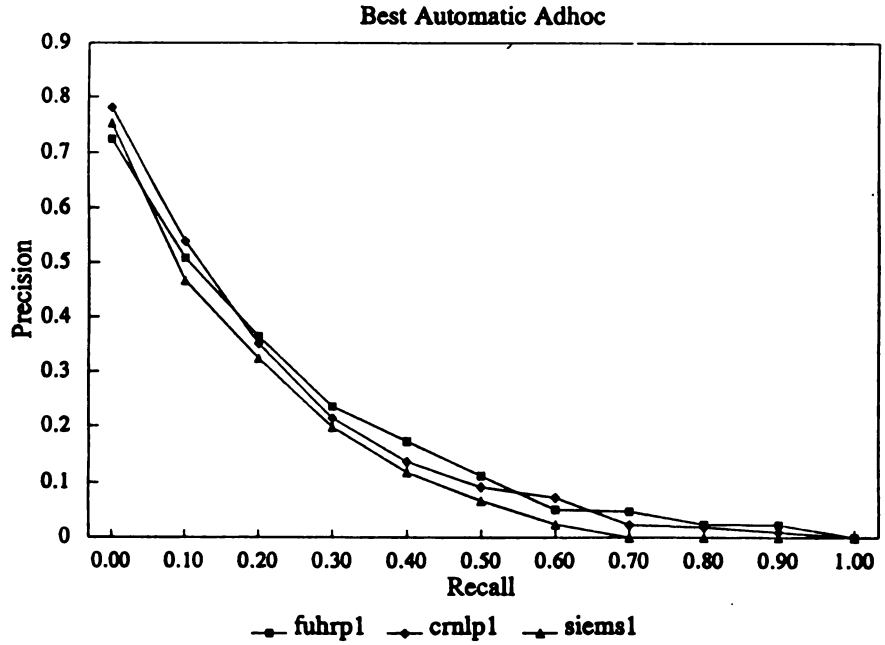

Figure 5. The Best Adhoc Results using Automatic Query Construction.

Figure 6 shows the recall/precision curve for the three TREC-1 runs with the highest 11-point averages using manual construction of queries. It should be noted that varying amounts of manual intervention were used, and this should be considered when comparing results. These curves show differences in that the "clartb" and "gecrd2" have initially a high precision, but lose this precision as recall increases, whereas the "cnqst2" method has a lower initial precision, but higher precision at the higher recall levels. This may be a function of the very different methods being used. The "clartb" system adds noun phrases found in likely relevant documents to improve the query terms taken from the topic (see Evans paper), whereas the "cnqst2" system uses more general thesaurus entries to expand the query (see Nelson paper). The "gecrd2" system uses a totally different approach of constructing elaborate Boolean pattern matchers (see Jacobs, Krupka & Rau paper).

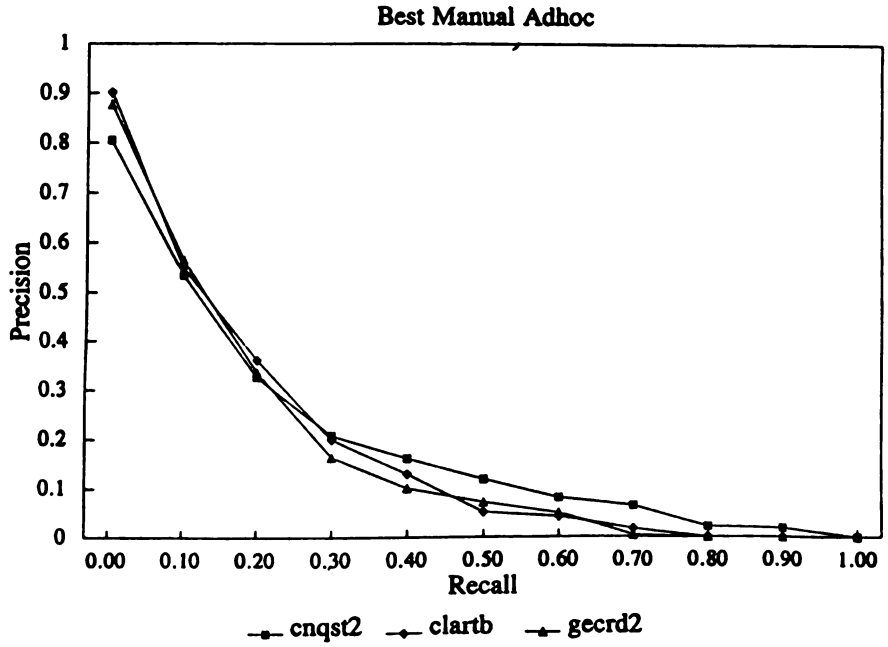

Figure 6. The Best Adhoc Results using Manual Query Construction.

It is useful to contrast the three methods of query construction. Figure 7 shows a comparison of five sets of results, two from automatic query construction, two using manual query construction, and the one relevance feedback run. It should be noted that there is relatively little difference between the results from automatic query construction versus manual query construction, although the relevance feedback results (citym2) were poor in this case. Figure 8 shows a histogram of the same information, but for all adhoc systems working with the full data set. In general it shows that the automatic query construction seems to work well for many systems, and that certainly it can be concluded that for TREC-1 the automatic construction of queries was as effective as the manual construction.

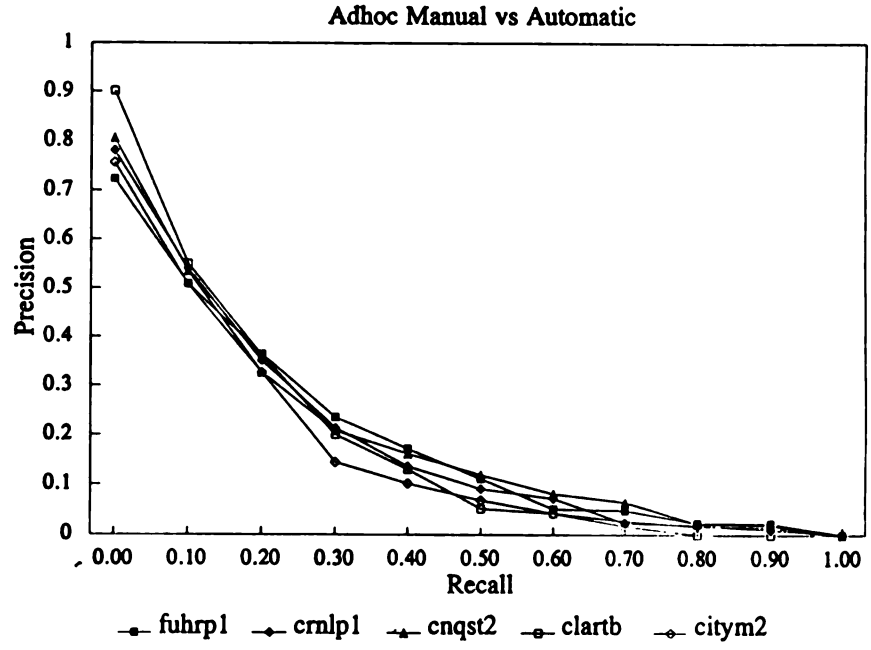

Figure 7. A Comparison of Adhoc Results using Different Query Construction Methods.

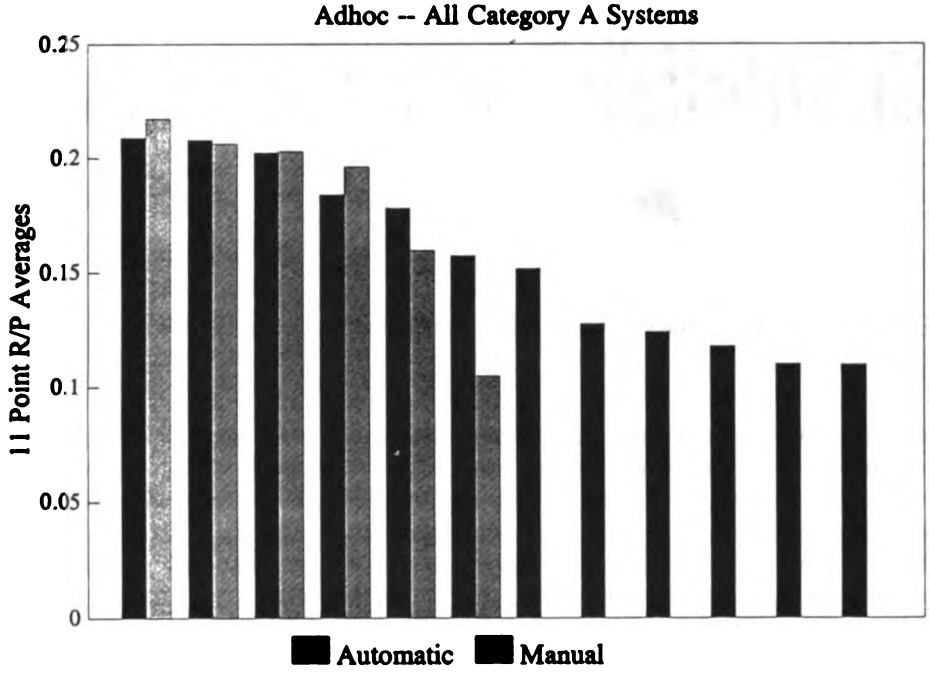

Figure 8. A Comparison of Adhoc Results using Automatic and Manual Query Construction.

Figure 9 shows the comparison of automatic and manual query construction on a per topic basis. It is interesting to note that, for the two systems shown, most topics show equal performance in terms of the percentage of relevant documents retrieved by 100 documents. Some topics, like topic 51, show much better manual performance, whereas other topics, like topic 69 show better automatic performance. This is somewhat different results from earlier comparisons of Boolean systems (usually manual indexing and manual query construction) versus the automatic systems such as the SMART system. In the Medlars study (Salton 1969) the manual (Boolean) systems seemed to do either very well or very poorly, whereas the automatic systems produced consistent "medium" results. The difference in the TREC task is likely that the topics are very long and complex, and sometimes are easy to express manually, but sometimes very difficult, whereas the automatic construction is hampered by the existence of difficult narratives. This is only a hypothesis and needs further investigation.

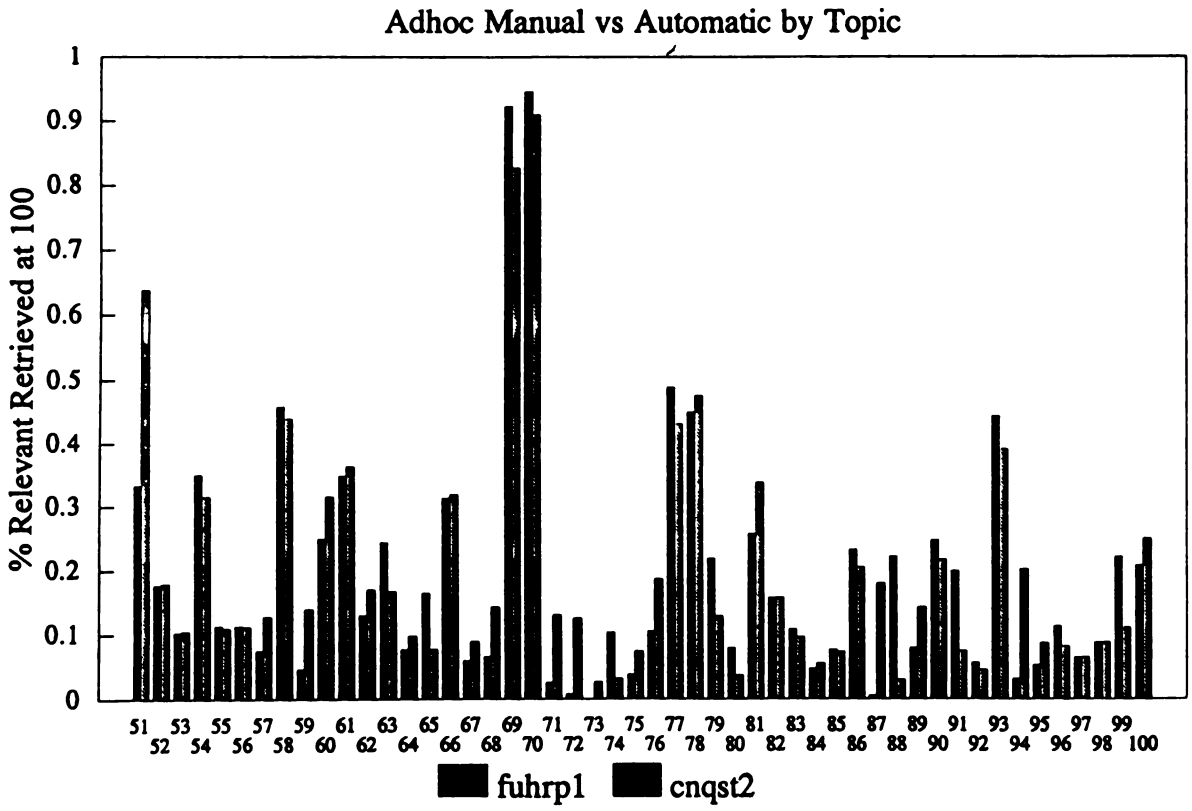

Figure 9. Adhoc Results using Automatic and Manual Query Construction.
on a Per Topic Basis

There were also some category B results for adhoc, and the best of these are shown in Figure 10, with results from the Cornell system run as a category B run to show some comparison. There is a wide spread in the curves here, with widely differing systems being shown. The "pircs4" results represent a very successful relevance feedback method, with the "pircs1" being an automatic query construction using the same system (see Kwok, Padadopoulos & Kwan paper). The "nyuir1" results come from a system using natural language techniques (see Strzalkowski paper).

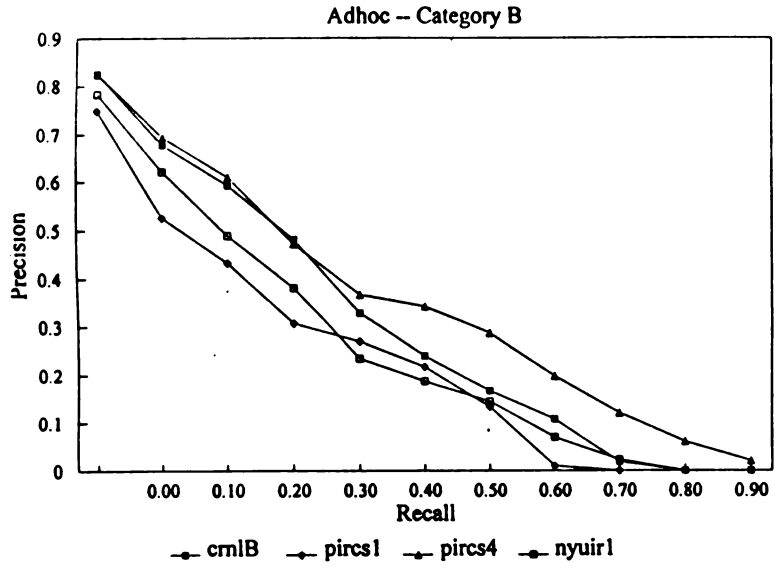

Figure 10. Adhoc Results for Category B.

5.3 Routing Results

There were 22 sets of results for routing evaluation, with 16 of them based on runs for the full data set. Note that all routing techniques suffered from the lack of sufficient and accurate training data, and therefore these results are even more preliminary than the adhoc results. Of the 16 systems using the full data set, 8 used automatic construction of queries, and 8 used manual construction. Figure 11 shows the recall/precision curve for the three TREC-1 runs with the highest 11-point averages using automatic construction of queries. Two of the curves, based on the use of the Cornell SMART system, show very different results. The "fuhra2" results came from using a probabilistically-based relevance feedback (see Fuhr & Buckley paper), whereas the "crnla2" results came from doing traditional relevance feedback methods using the vector space model (see Buckley, Salton & Allan paper), The "cityr1" results also came from using traditional relevance feedback, but using a different probabilistic model and term weighting (see Robertson, Walker, Hancock-Beaulieu, Gull & Lau paper). The "cpgcn2" system used filtering methods rather than more traditional information retrieval methods to achieve results similar to the feedback results (see Jones, Leung, and Pape paper).

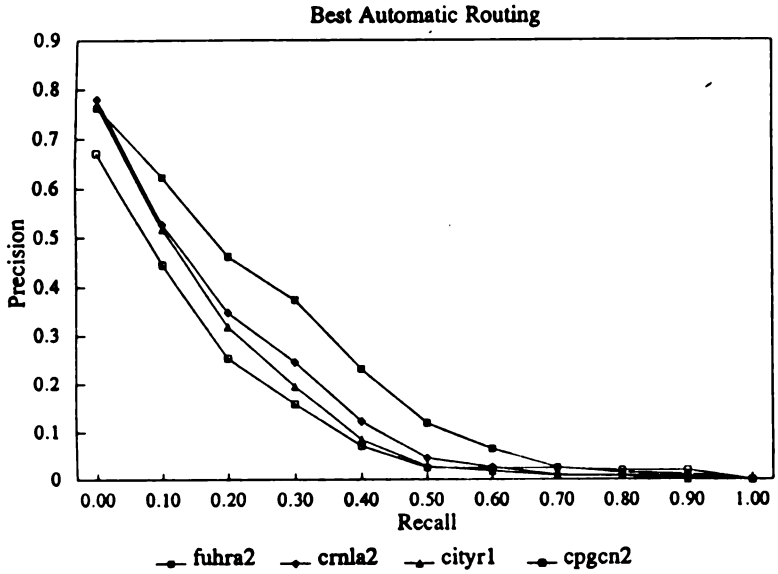

Figure 11. The Best Routing Results using Automatic Query Construction.

Figure 12 shows the recall/precision curve for the three TREC-1 runs with the highest 11-point averages using manual construction of queries. The systems used manually-built filters, with the "clartb" and "gecrd2" results done similarly to their corresponding adhoc systems, but using the sample relevant documents as input to the filter-building process. The "paraz1" system used manually-constructed filters based on clusters of interesting terms (see Zimmerman paper). The "cpghc2" group hand-crafted these queries as a contrast to their automatic pattern filtering methods (see Jones, Leung & Pape).

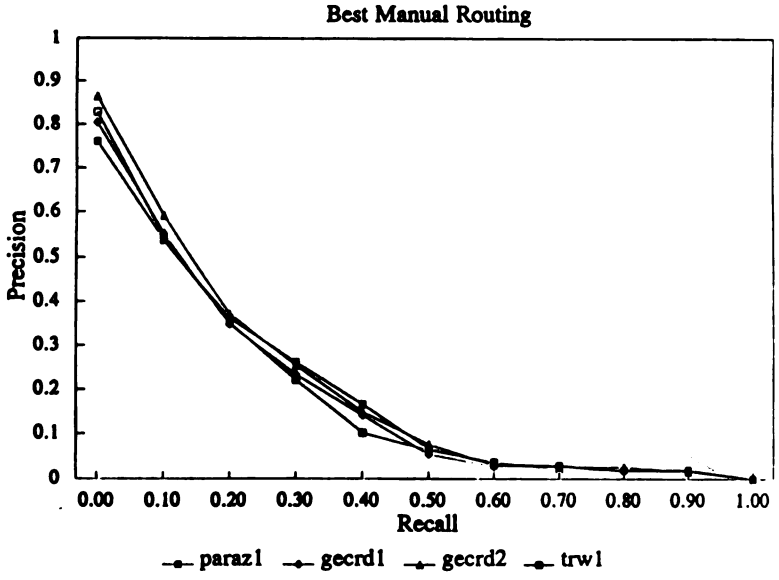

Figure 12. The Best Routing Results using Manual Query Construction.

Again it is useful to contrast the methods of query construction. Figure 13 shows a comparison of four sets of results, two from automatic query construction and two using manual query construction. Here, unlike the adhoc results, the automatic query building seems to be clearly superior, with the "fuhr1" results having higher performance throughout the significant part of the recall/precision curve.

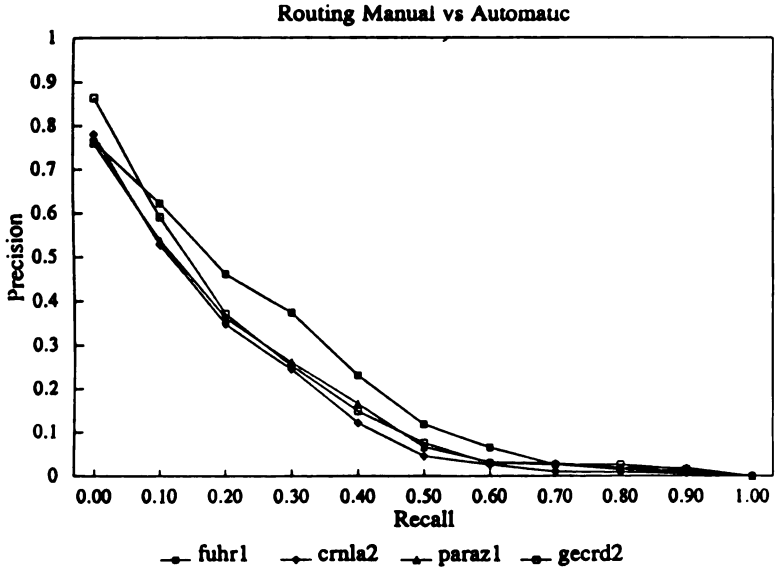

Figure 13. A Comparison of Routing Results using Different Query Construction Methods.

There were also some category B results for routing, and the best of these are shown in Figure 14. Again there is a much wider spread in the curves here, and widely differing systems being shown. The "pircs1" and "pircs2" results are correspondingly automatic and manually constructed queries using relevance feedback learning from the training sample (see Kwok, Padadopoulos & Kwan paper). The "fairs1" system uses a combination of different term weighting methods (see Chang, Dediu, Assam & Du paper). The "nyuir1" results come from a system using natural languages techniques and probably reflect the short amount of time available to construct this very complicated system (see Strzalkowski paper).

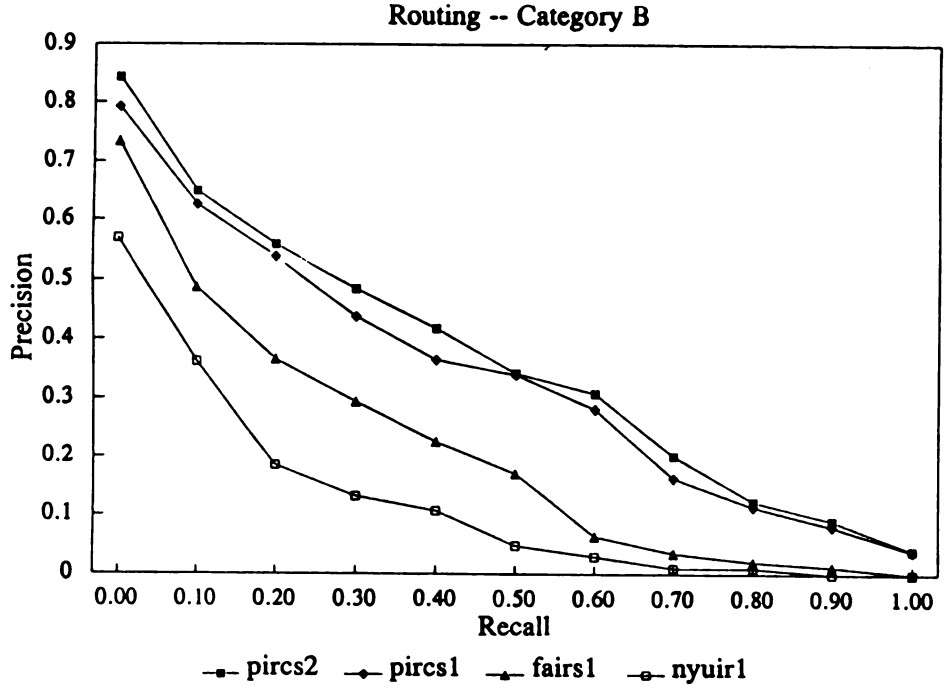

Figure 14. Routing Results for Category B.

5.4 Summary

The TREC-1 conference demonstrated a wide range of different approaches to the retrieval of text from large document collections. Because of the preliminary nature of the results, very little can be said about which techniques seem to perform the best. It was clear that the simple systems did the task well, but it is too early to pass judgment on the more complicated systems. The automatic construction of queries from the topics did as well as, or better than, manual construction of queries, and this is encouraging for groups supporting the use of simple natural language interfaces for retrieval systems.

There will be a second TREC conference in 1993, and all the systems that participated in TREC-1 will be back, along with additional groups. The results from this second conference should better identify the more promising retrieval techniques.

REFERENCES

Cleverdon C.W., (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems.* College of Aeronautics, Cranfield, England, 1962.

Cleverdon C.W., Mills, J. and Keen E.M. (1966). *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results*. Aslib Cranfield Research Project, Cranfield, England, 1966.

Harman D. (1993). The DARPA TIPSTER Project. *SIGIR Forum*, 26(2), 26-28.

Katzer J., McGill M.J., Tessier J.A., Frakes W., and DasGupta P. (1982). A Study of the Overlap among Document Representations. *Information Technology: Research and Development*, 1(2), 261-274.

Fox E. (1983). Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts. *Technical Report TR 83-561*, Cornell University: Computing Science Department.

Salton G. (1969). A Comparison Between Manual and Automatic Indexing Methods. *American Documentation*, 20(1).

Salton G. and McGill M. (1983). *Introduction to Modern Information Retrieval*. New York, NY.: McGraw-Hill.

Sparck Jones K. (1981). *Information Retrieval Experiment*. London, England: Butterworths.

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. and Webster (1979). *Research in Relevance Weighting*, British Library Research and Development Report 5553, Computer Laboratory, University of Cambridge.

Swets J. (1969). Effectiveness of Information Retrieval Methods. *American Documentation*, 20(1).