OVERVIEW OF TREC-1

Donna Harman

National Institute of Standards and Technology
Gaithersburg, Md. 20899

ABSTRACT

The first Text REtrieval Conference (TREC-1) was held in early November 1992 and was attended by about 100 people working in the 25 participating groups. The goal of the conference was to bring research groups together to discuss their work on a new large test collection. There was a large variety of retrieval techniques reported on, including methods using automatic thesaurii, sophisticated term weighting, natural language techniques, relevance feedback, and advanced pattern matching. As results had been run through a common evaluation package, groups were able to compare the effectiveness of different techniques, and discuss how differences among the systems affected performance.

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 [1], and has continued with over 30 years of experimentation with the retrieval engines themselves. The Cranfield II studies [2] 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 [3], and the NPL collection [4].

In the thirty 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 [5] 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 ten 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 thesaurii 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) is 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.

The following description was excerpted from a more lengthy overview published in the conference proceedings [6]. The full proceedings also contain papers by all participants and results for all systems.

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 "ad-hoc" 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. It is assumed that potential 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 ad-hoc 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 (Figure 1).

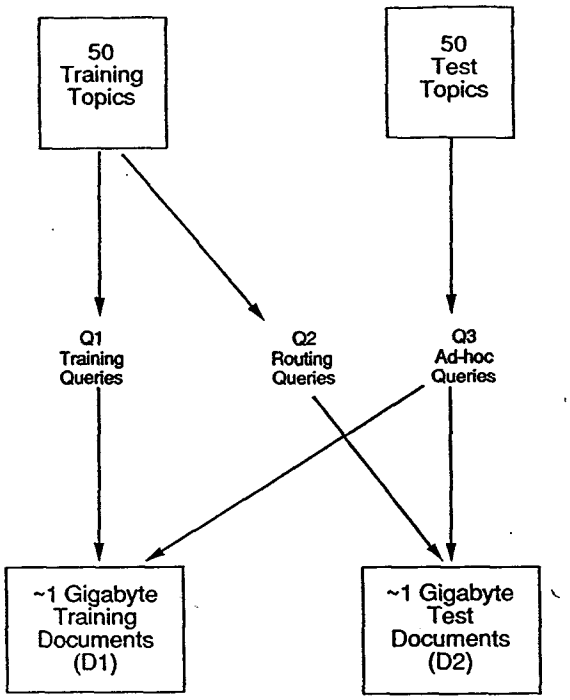

Figure 1 -- The TREC Task

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, 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 ad-hoc 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 queries could be constructed using one of three alternative methods. They could be constructed automatically from the topics, with no human intervention. Alternatively they could be constructed manually from the topic, but with no "retries" after looking at the results. The third method allowed "retries", but under constrained conditions.

2.2 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 twenty 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 chose 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

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

3.1 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 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 tasks. 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.

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.

3.2 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 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 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>

3.3 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. Relevance judgments were made using a sampling method, with the sample constructed by taking the top 100 documents retrieved by each system for a given topic and merging them into a pool for relevance assessment. This sampling, known as pooling, proved to be an effective method. There was little overlap among the 25 systems in their retrieved documents. For example, out of a maximum of 3300 unique documents (33 runs times 100 documents), over one-third were actually unique. This means that the different systems were finding different documents as likely relevant documents for a topic. 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 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. 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 (with a median of about 250 documents).

4. PRELIMINARY RESULTS

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 the proceedings. The results of the TREC-1 conference can 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. Additionally groups were working blindly as to what constitutes a relevant document. There were no reliable relevance judgments for training, and the use of the structured topics was completely new. 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.

However there were some clear trends that emerged. Automatic construction of queries proved to be as effective as manual construction of queries. Figure 2 shows a comparison of four sets of results, two using automatic query construction and two using manual query construction, and it can be seen that there is relatively little difference between the results.

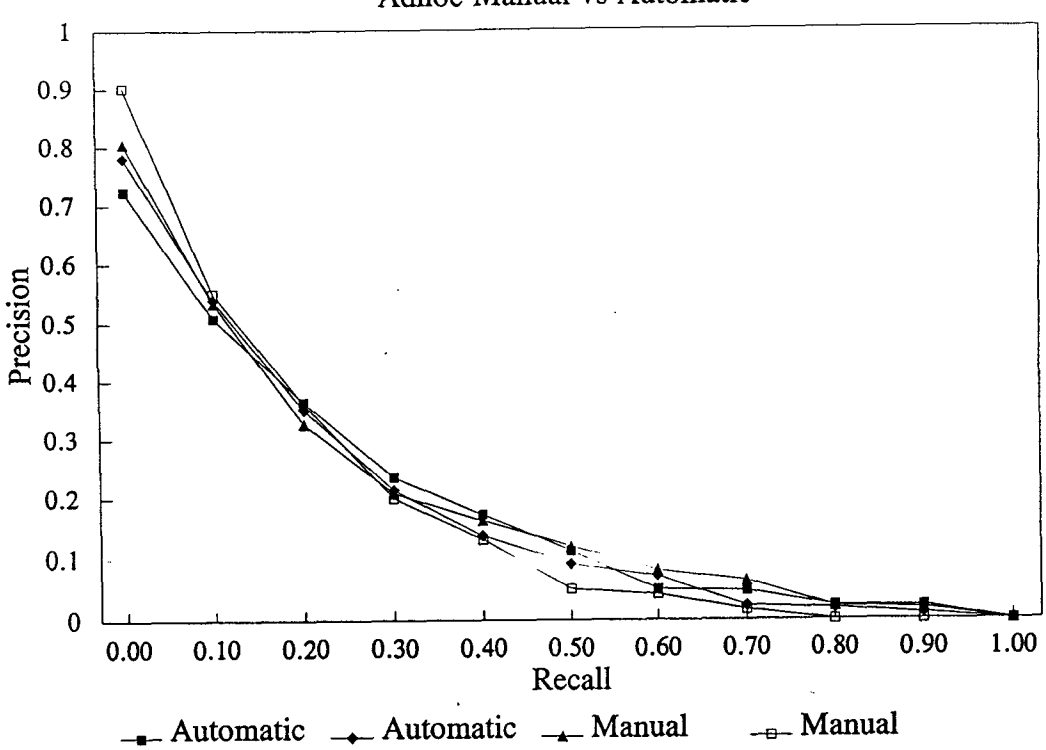

Figure 2 -- A Comparison of Adhoc Results using Different Query Construction Methods

The two automatic systems shown used basically all the terms in the topic as query terms, and relied on automatic term weighting and sophisticated ranking algorithms for performance. The manual systems also used sophisticated term weighting and algorithms, but manually selected which terms to include in a query.

Several minor trends were also noticeable. Systems that worked with subdocuments, or used local term context to improve term weighting, seemed particularly successful in handling the longer documents in TREC. More systems may investigate this approach in TREC-2. Also systems that attempted to expand a topic beyond its original terms (either manually or automatically) seemed to do well, although it was often hard to properly control this expansion (particularly for automatically expanded queries). These trends may continue in TREC-2 and it is expected that clearer trends will emerge as groups have more time to work at this new task.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Sparck Jones K. and C. Webster (1979). *Research in Relevance Weighting*, British Library Research and Development Report 5553, Computer Laboratory, University of Cambridge.

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

[6] Harman D. "The First Text REtrieval Conference (TREC1)." *National Institute of Standards and Technology Special Publication 500-207*, Gaithersburg, Md. 20899 (in press, available in May 1993).