OVERVIEW OF THE SECOND TEXT RETRIEVAL CONFERENCE (TREC-2)

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). This conference, co-sponsored by ARPA and NIST, brought together information retrieval researchers to discuss their system results on the new TIPSTER test collection. This was the first time that such groups had ever compared results on the same data using the same evaluation methods, and represented a breakthrough in cross-system evaluation in information retrieval. It was also the first time that most of these groups had tackled such a large test collection and required a major effort by all groups to scale up their retrieval techniques.

Since TREC is designed to evaluate system performance both in a routing (filtering or profiling) mode, and in an adhoc mode, both functions were tested. The test design was based on traditional information retrieval models, involving documents, "user" questions, and the "right answers" (Harman 1994a). Participants were first sent two disks of documents (about 2 gigabytes of data) and a training set of 100 questions or topics. They were also sent lists of documents in the two disks that were considered the "right answers" or relevant documents for each of the 100 topics. The participants were asked to train their systems on this data, and at some point to signal their readiness for testing by submitting their system queries for a specific fifty of the topics. The routing test consisted of each group running new test documents against those 50 queries. The adhoc test consisted of running a new set of 50 topics against the old document set (the original 2 disks). In each case, the results of the retrieval systems were submitted to NIST for evaluation.

The documents in the test collection are from various types of text, covering different writing styles and different information domains. They include information from the Wall Street Journal, the San Jose Mercury News, the AP Newswire, and artcles from the Computer Select disks. The documents were uniformly formatted into an SGML-like structure for easy handling by the TREC participants.

The topics used in the test collection are in the form of "user need" statements rather than more traditional queries. They are 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 is diverse, some consideration was given to the documents to be searched.

The relevance judgments or "right answers" were made using a sampling method, with the sample constructed by taking the top 100 documents retrieved by each participating system for a given topic and merging them into a pool for manual relevance assessment. This is a valid sampling method since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first. All systems were then evaluated against the common set of relevant documents, i.e. the total number of relevant documents found by all the systems combined.

How well did the systems do with this test collection? Whereas the TREC-1 conference demonstrated a wide range of different approaches to the retrieval of text from large document collections, the results could be viewed only as very preliminary. Not only were the deadlines for results were very tight, but 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. Still, the results showed that the systems did the task well, and that automatic construction of queries from the topics did as well as, or better than, manual construction of queries.

The second TREC conference (TREC-2) occurred in August of 1993, less than 10 months after the first conference. In addition to most of the TREC-1 groups, nine new groups took part, bringing the total number of participating groups to 31.

Advanced Decision Systems	Bellcore
Carnegie Mellon University	CITRI, Australia
City University, London	Conquest, Inc.
Cornell University	Dalhousie University
Environment Research Institute of Michigan	GE Research and Development Center
HNC Inc.	Institute for Decision Systems Research
Mead Data Central	New York University
PRC, Inc.	Queens College
Rutgers University	Siemens Corporate Research Inc.
Swiss Federal Institute of Technology (ETH)	Syracuse University
Systems Environment Corporation	Thinking Machines Corporation
TRW Systems Development Division	Universitaet Dortmund, Germany
University of California - Berkeley	University of California - UCLA
University of Central Florida	University of Illinois at Chicago
University of Massachusetts at Amherst	VPI&SU (Virginia Tech)
Verity Inc.	

Table 1: TREC-2 Participants (14 companies, 17 universities)

2. TREC-2 RESULTS

2.1 Introduction

In general the TREC-2 results showed significant improvements over the TREC-1 results. Many of the original TREC-1 groups were able to "complete" their system rebuilding and tuning tasks. The results for TREC-2 therefore can be viewed as the "best first-pass" that most groups can accomplish on this large amount of data. The adhoc results in particular represent baseline results from the scaling-up of current algorithms to large test collections. The better systems produced similar results, results that are comparable to those seen using these algorithms on smaller test collections.

The routing results showed even more improvement over TREC-1 routing results. Some of this improvement was due to the availability of large numbers of accurate relevance judgments for training (unlike TREC-1), but most of the improvements came from new research by participating groups into the best ways of using the training data.

All references in this section are papers in the TREC-2 proceedings (Harman 1994b).

2.2 Adhoc Results

The adhoc evaluation used new topics (101-150) against the two disks of training documents (disks 1 and 2). There were 44 sets of results for adhoc evaluation in TREC-2, with 32 of them based on runs for the full data set. Of these, 23 used automatic construction of queries, 9 used manual construction, and 2 used feedback.

Figure 1 shows the recall/precision curves for the six TREC-2 groups with the highest non-interpolated average precision using automatic construction of queries. The results marked "INQ001" are the INQUERY system from the University of Massachusetts (see Croft, Callan & Broglio paper). This system uses probabilistic term weighting and a probabilistic inference net to combine various topic and document features. The results marked "dortQ2", "Brkly3" and "crnlL2" are all based on the use of the Cornell SMART system, but with important variations. The "crnlL2" run is the basic SMART system from Cornell University (see Buckley, Allan & Salton paper), but using less than optimal term weightings (by mistake). The "dortQ2" results from the University of Dortmund come from using polynomial regression on the training data to find weights for various pre-set term features (see Fuhr, Pfeifer, Bremkamp, Pollmann & Buckley paper). The "Brkly3" results from the University of California at Berkeley come from performing logistic regression analysis to learn optimal weighting for various term frequency measures (see Cooper, Chen & Gey paper). The "CLARTA" system from the CLARIT Corporation expands each topic with noun phrases found in a thesaurus that is automatically generated for each topic (see Evans & Lefferts paper). The "lsiasm" results are from Bellcore (see Dumais paper). This group uses latent semantic indexing to create much larger vectors than the more traditional vector-space models such as SMART. The run marked "lsiasm" represents only the base SMART pre-processing results, however. Due to processing errors the "improved" LSI run produced unexpectedly poor results.

Figure 2 shows the recall/precision curve for the six TREC-2 groups with the highest non-interpolated average precision using manual construction of queries. It should be noted that varying amounts of manual intervention were used. The results marked "INQ002", "siems2", and

"CLARTM" are automatically-generated queries with manual modifications. The "INQ002" results reflect various manual modifications made to the "INQ001" queries, with those modifications guided by strict rules. The "siems2" results from Siemens Corporate Research, Inc. (see Voorhees paper) are based on the use of the Cornell SMART system, but with the topics manually modified (the "not" phases removed). These results were meant to be the base run for improvements using WordNet, but the improvements did not materialize. The "CLARTM" results represent manual weighting of the query terms, as opposed to the automatic weighting of the terms that was used in "CLARTA". The results marked "Vtcms2", "CnQst2", and "TOPIC2" are produced from queries constructed completely manually. The "Vtcms2" results are from Virginia Tech (see Fox & Shaw paper) and show the effects of combining the results from SMART vector-space queries with the results from manually-constructed soft Boolean P-Norm type queries. The "CnQst2" results, from ConQuest Software (see Nelson paper), use a very large general-purpose semantic net to aid in constructing better queries from the topics, along with sophisticated morphological analysis of the topics. The results marked "TOPIC2" are from the TOPIC system by Verity Corp. (see Lehman & Reid paper) and reflect the use of an expert system working off specially-constructed knowledge bases to improve performance.

Several comments can be made with respect to these adhoc results. First, the better results (most of the automatic results and the three top manual results) are very similar and it is unlikely that there is any statistical differences between them. There is clearly no "best" method, and the fact that these systems have very different approaches to retrieval, including different term weighting schemes, different query construction methods, and different similarity match methods implies that there is much more to be learned about effective retrieval techniques. Additionally, whereas the averages for the systems may be similar, the systems do better on different topics and retrieve different subsets of the relevant documents.

A second point that should be made is that the automatic query construction methods continue to perform as well as the manual construction methods. Two groups (the INQUERY system and the CLARIT system) did explicit comparision of manually-modified queries vs those that were not modified and concluded that manual modification provided no benefits. The three sets of results based on completely manually-generated queries had even poorer performance than the manually-modified queries. Note that this result is specific to the very rich TREC topics; it is not clear that this will hold for the short topics normally seen in other retrieval environments.

As a final point, it should be noted that these adhoc results represent significant improvements over the results from TREC-1. Figure 5 (after the routing results) shows a comparison of results for a typical system in TREC-1 and TREC-2. Some of this improvement is due to improved evaluation, but the difference between the curve marked "TREC-1" and the curve marked "TREC-2 looking at top 200 only" shows significant performance improvement. Whereas this improvement could represent a difference in topics (the TREC-1 curve is for topics 51-100 and the TREC-2 curves are for topics 101-150), the TREC-2 topics are generally felt to be more difficult and therefore this improvement is likely to be an understatement of the actual improvements.

Very few groups worked with less than the full document collection. The system from New York University (see Strzalkowski & Carballo paper) reflects a very intensive use of natural language processing (NLP) techniques, including a parse of the documents to help locate syntactic phrases, context-sensitive expansion of the queries, and other NLP improvements on statistical techniques. In interests of space this graph is not shown; please refer to the paper by this group in this proceedings.

2.3 Routing Results

The routing evaluation used a subset of the training topics (topics 51-100 were used) against the new disk of test documents (disk 3). There were 40 sets of results for routing evaluation, with 32 of them based on runs for the full data set. Of the 32 systems using the full data set, 23 used automatic construction of queries, and 9 used manual construction.

Figure 3 shows the recall/precision curves for the six TREC-2 groups with the highest non-interpolated average precision using automatic construction of the routing queries. Again three systems are based on the Cornell SMART system. The plot marked "crnlC1" is the actual SMART system, using the basic Rocchio relevance feedback algorithms, and adding many terms (up to 500) from the relevant training documents to the terms in the topic. The "dortP1" results come from using a probabilistically-based relevance feedback instead of the vector-space algorithm, and adding only 20 terms from the relevant documents to each query. These two systems have the best routing results. The "Brkly5" system uses logistic regression on both the general frequency variables used in their adhoc approach and on the query-specific relevance data available for training with the routing topics. The results marked "cityr2" are from City University, London (see Robertson, Walker, Jones, Hancock-Beaulieu & Gafford paper). This group automatically selected variable numbers of terms (10-25) from the training documents for each topic (the topics themselves were not used as term sources), and then used traditional probabilistic reweighting to weight these terms. The "INQ003" results also use probabilistic reweighting, but use the topic terms, expanded by 30 new terms per topic from the training documents. The results marked "lsir2" are more latent semantic indexing results from Bellcore. This run was made by creating a filter of the singular-value decomposition vector sum or centroid of all relevant documents for a topic (and ignoring the topic itself).

Figure 4 shows the recall/precision curves for the six TREC-2 groups with the highest non-interpolated average precision using manual construction of the routing queries. The results marked "INQ004" are from the INQRY system using an inferential combination of the "INQ003" queries and manually modified queries created from the topic. The "trw2" results represent an adaptation of the TRW Fast Data Finder pattern matching system to allow use of term weighting (see Mettler paper). The queries were manually constructed and the term weighting was learned from the training data. The "gecrd1" results from GE Research and Development Center (see Jacobs paper) also come from manually-constructed queries, but using a general-purpose lexicon and the training data to suggest input to the Boolean pattern matcher. The results marked "CLARTM" are similar to the "CLARTM" adhoc results except that the training documents were used as the source for thesaurus building, as opposed to using the top set of retrieved documents. The "rutcombx" results from Rutgers University (see Belkin, Kantor, Cool & Quatrain paper) come from combining 5 sets of manually-generated Boolean queries to optimize performance for each topic. The results marked "TOPIC2" are from the TOPIC system and reflect the use of an expert system working off specially-constructed knowledge bases to improve performance.

As was the case with the adhoc topics, the automatic query construction methods continue to perform as well as, or in this case, better than the manual construction methods. A comparision of the two INQRY runs illustrates this point and shows that all six results with manually-generated queries perform worse than the six runs with automatically-generated queries. The availability of the training data allows an automatic tuning of the queries that would be difficult to duplicate manually without extensive analysis.

Unlike the adhoc results, there are two runs ("crnlC1" and "dortP1") that are clearly better than the others, with a significant difference between the "crnlC1" results and the "dortP1" results and also significant differences between these results and the rest of the automatically-generated query results. In particular the use of so many terms (up to 500) for query expansion by the Cornell group was one of the most interesting findings in TREC-2 and represents a departure from past results (see Buckley, Allan, & Salton paper for more on this).

As a final point, it should be noted that the routing results also represent significant improvements over the results from TREC-1. Figure 6 shows a comparison of results for a typical system in TREC-1 and TREC-2. Some of this improvement is due to improved evaluation, but the difference between the curve marked "TREC-1" and the curve marked "TREC-2 looking at top 200 only" shows significant performance improvement. There is more improvement for the routing results than for the adhoc results due to better training data (mostly non-existent for TREC-1) and to major efforts by many groups in new routing algorithm experiments.

3. SUMMARY

The TREC-2 conference demonstrated a wide range of different approaches to the retrieval of text from large document collections. There was significant improvement in retrieval performance over that seen in TREC-1, especially in the routing task. The availability of large amounts of training data for routing allowed extensive experimentation in the best use of that data, and many different approaches were tried in TREC-2. The automatic construction of queries from the topics continued to do 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. The conference itself continued to provide an open forum for exchange of results, and the increased participation by commercial groups will speed the transfer of TREC algorithms into readily-available software products.

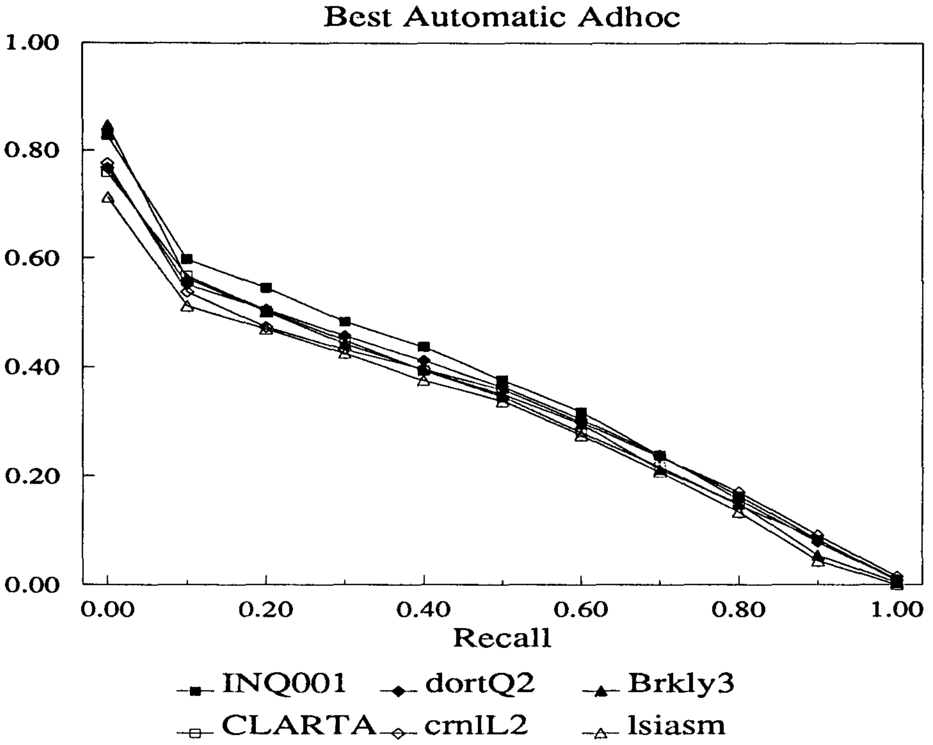
There is a TREC-3 planned for November 1994, with most of the TREC-2 participants returning, and a current roster of over 55 groups participating.

4. 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. (1994a). Data Preparation. In: Merchant R. (Ed.).*The Proceedings of the TIPSTER Text Program - Phase I.* San Mateo, California: Morgan Kaufmann Publishing Co., 1994.

Harman D. (Ed.). (1994b). *The Second Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology Special Publication 500-215, Gaithersburg, Md. 20899.



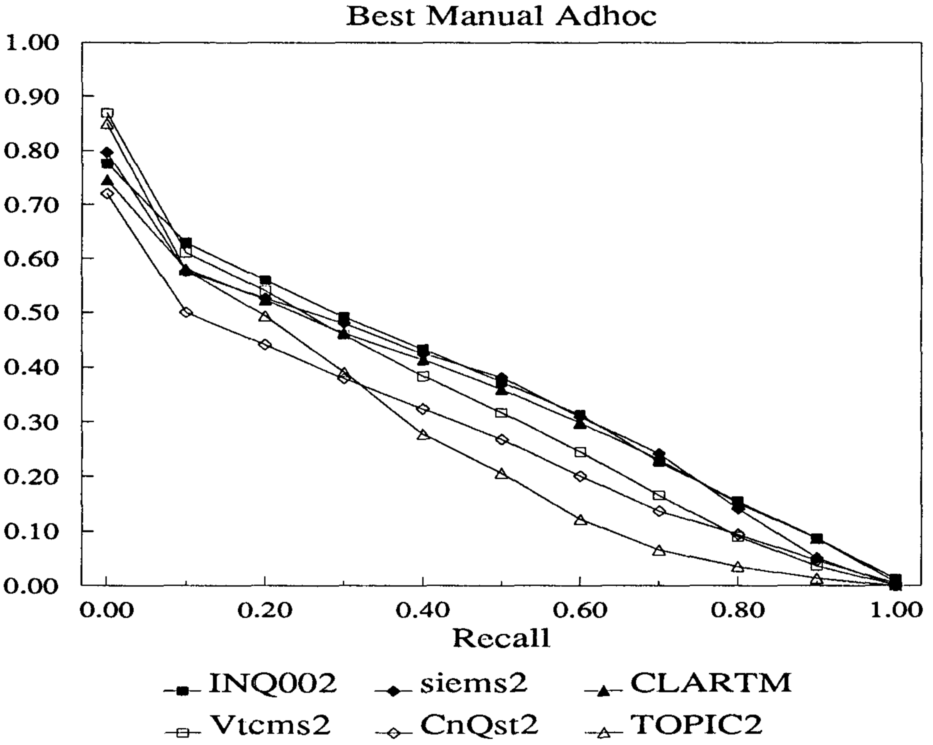
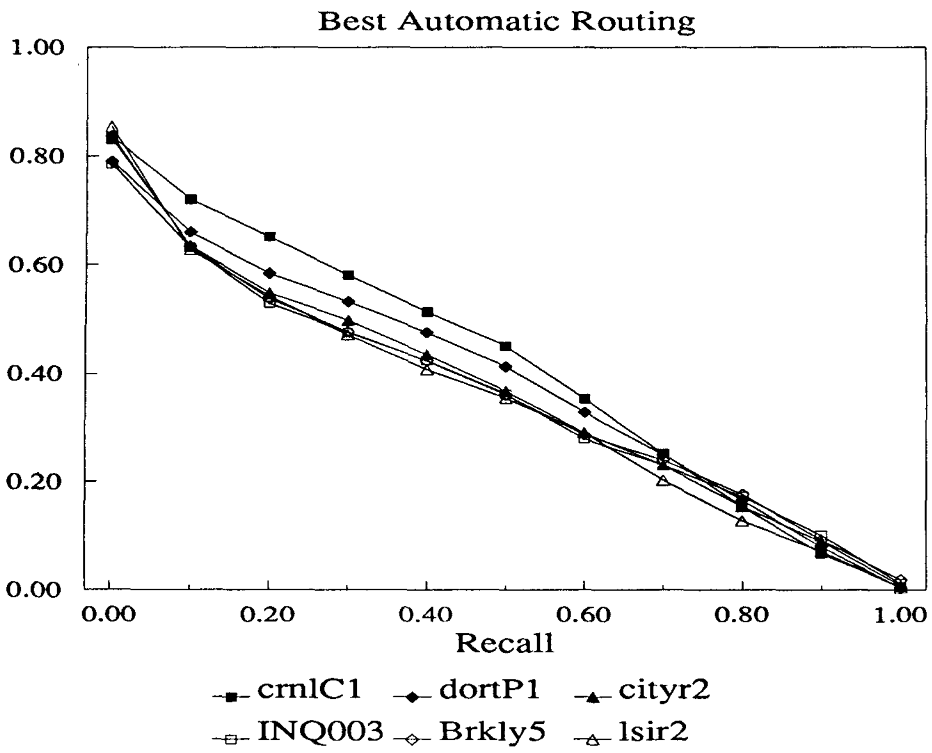


Figure 1 -- Best Automatic Adhoc Results
Figurc 2 -- Best Manual Adhoc Rcsults



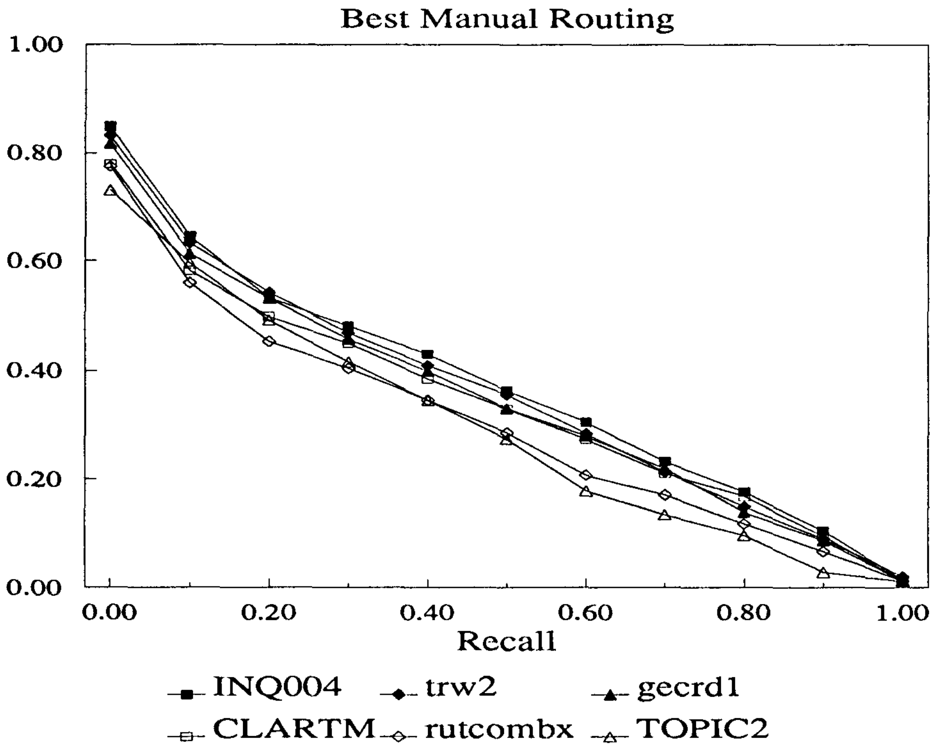
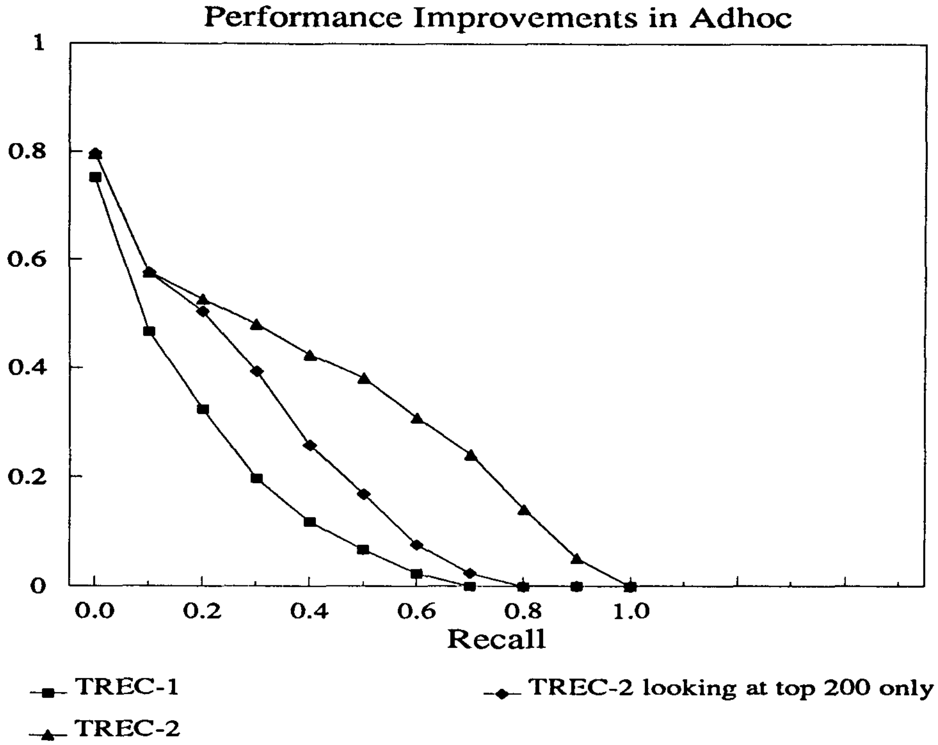


Figure 3 -- Best Automatic Routing Results
Figure 4 -- Best Manual Routing Results



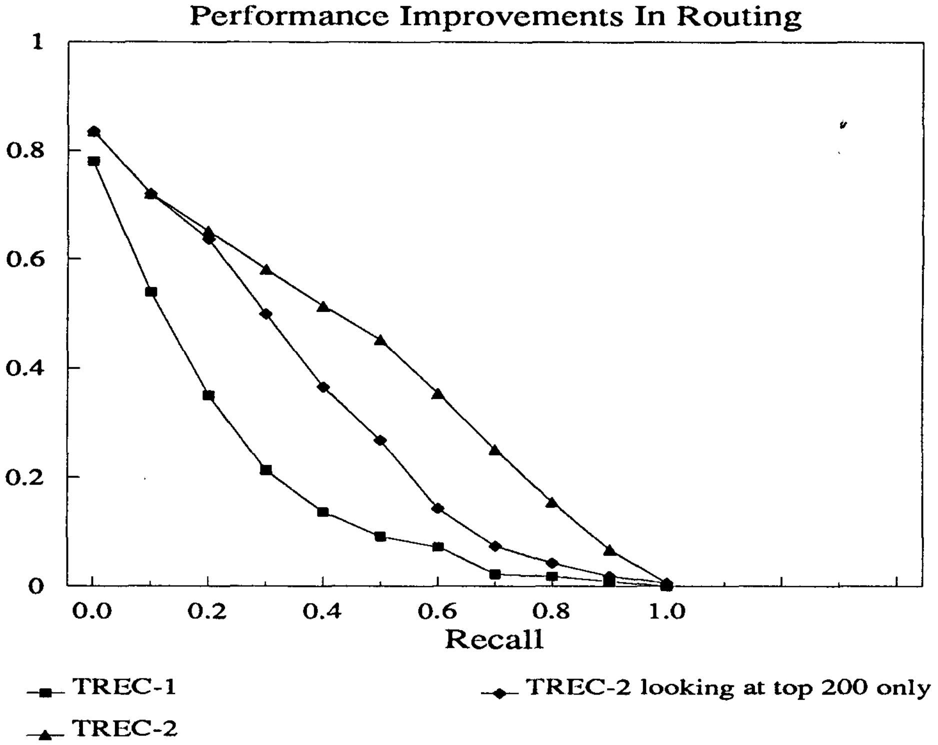


Figure 5 -- Typical Improvements in Adhoc Results
Figure 6 -- Typical Improvements in Routing Results