DOCUMENT DETECTION DATA PREPARATION

Donna Harman

National Institute of Standards and Technology
Gaithersburg, MD. 20899

1. THE ENGLISH TEST COLLECTION

1.1 Introduction

Critical to the success of TIPSTER was the creation of the test collection. Like most traditional retrieval collections, there are three distinct parts to this collection--the documents, the queries or topics, and the relevance judgments or "right answers". It was important to match all three parts of the collection to the TIPSTER application.

The document collection needed to reflect the corpus imagined to be seen by analysts. This meant that a very large collection was needed to test the scaling of the algorithms, including documents from many different domains to test the domain independence of the algorithms. Additionally the documents selected needed to mirror the different types of documents used in the TIPSTER application. Specifically they had to have a varied length, a varied writing style, a varied level of editing and a varied vocabulary. As a final requirement, the documents had to cover different timeframes to show the effects of document date on the routing task.

The topics for the TIPSTER collection were also designed to model some of the needs of analysts. It was assumed that the typical user of these retrieval systems was a dedicated searcher, not a novice searcher, and that the model for the application was one needing both the monitoring of data streams for information on specific topics (routing), and the ability to do adhoc searches on archived data for new topics. It was also 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. The topics therefore were created to be very specific, but included both broad and narrow searching needs. Many of the topics were created to test performance on specific types of searches.

The relevance assessments were made by retired analysts who were asked to view the task as if they were addressing a real information need. The narrative section of the topic (described in more detail later) contained a clear definition of what made a document relevant, and the assessors used this section as the definition of the information need. Each topic was judged by a single assessor so that all documents screened would reflect the same user's interpretation of topic.

1.2 The Documents

The documents came from many different sources. These sources were selected not only because of their suitability to the TIPSTER task, but also because of their availability. The data was provided by the University of Pennsylvania, initially as part of the ACL/DCI text initiative, but later as part of the Linguistic Data Consortium effort. A successful pattern for data source selection was established for the first disk and was followed for disks 2 and 3. First, two sets of documents were obtained that contained articles from all domains. The first set of articles was from a newspaper (the Wall Street Journal in disks 1 and 2, the San Jose Mercury News in disk 3), and the second set of articles was from a newswire (AP in all three disks). In addition to covering all domains, these two sets provide a strong contrast in their format, style, and accuracy of editing, and were both readily available. The third set of documents was selected to cover more deeply a particular domain. Partially because of availability, the particular set used was a subset of the Computer Select disks, from Ziff-Davis publishing. These documents cover the wide domain of computers and computer technology, but include many different sources of actual documents. This creates a set of documents in a single (broad) domain, but having a range of formatting and writing styles. The final set of documents (or final two sets in the first disk) were selected less for content than for length of articles. Since the documents in the first three sets were of medium length, and of fairly uniform length, the final set of documents was picked to be especially long, and of non-uniform length. Documents from the Federal Register were used for the first two disks, and some U.S. Patents were used on disk 3. Additionally the first disk contained some very short abstracts from the Department of Energy.

Because most of this material is copyrighted, all users of this data were required to sign a detailed agreement in order to protect the copyrighted source material.

The following shows the actual contents of each disk.

- 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)
- Disk 3
    - SJMN -- San Jose Mercury News (1991)
    - AP -- AP Newswire (1990)
    - ZIFF -- Information from Computer Select disks (Ziff-Davis Publishing)
    - PAT -- U.S. Patents (1993)

All documents were originally received at the University of Pennsylvania in various print-tape formats. These formats were converted to an SGML-like structure and sent to NIST. At NIST the documents were assigned unique document identifiers and the formats were more standardized. The following example shows an abbreviated version of a typical document.

```
<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 have beginning and end markers, and a unique DOCNO id field. Additionally other fields taken from the initial data appear, but these vary widely across the different sources. The documents have differing amounts of errors, which were not checked or corrected. Not only would this have been an impossible task, but the errors in the data provide a better simulation of the TIPSTER task. Errors in missing document separators or bad document numbers were screened out, although a few were missed and later reported as errors.

Table 1 shows some basic document collection statistics. 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 covers many domains, resulting in many specific technical terms. The typical distribution of terms in the collections in general corresponds to Zipf's law [1] in that about half the total number of unique terms only appear once. This is least applicable to the newspaper data, possibly because the vocabulary used is more controlled than in the other collections. The newspaper data also has the highest number of occurrences of terms for those terms appearing more than once.

Subset of collection	WSJ SJMN	AP	ZIFF	FR PAT	DOE
Size of collection (megabytes)					
(disk 1)	295	266	251	258	190
(disk 2)	255	248	188	211	
(disk 3)	315	248	358	251	
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	
(disk 3)	90,257	78,325	161,021	6,711	
Median number of terms per record					
(disk 1)	182	353	181	313	82
(disk 2)	218	346	167	315	
(disk 3)	279	358	119	2896	
Average number of terms per record					
(disk 1)	329	375	412	1017	89
(disk 2)	377	370	394	1073	
(disk 3)	337	379	263	3543	

Table 1: Document Statistics

Subset of collection	WSJ SJMN	AP	ZIFF PAT	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	
(disk 3)	179,490	190,278	212,729	156,077	
Occurring once					
(disk 1)	64,656	89,627	85,992	58,677	95,782
(disk 2)	64,844	83,019	72,053	54,823	
(disk 3)	73,064	86,976	106,857	90,128	
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	
(disk 3)	106,426	103,302	105,872	65,949	
Average number of occurrences > 1					
(disk 1)	199	174	165	106	159
(disk 2)	178	169	139	91	
(disk 3)	168	169	205	63	

Table 2: Dictionary Statistics

1.3 The Topics

Traditional information retrieval test collections have typically included sentence-length queries. These queries are usually automatically transformed into a machine version for searching, with minimal changes. In designing the TIPSTER task, there was a conscious decision made to provide "user need" statements rather than the 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. Topics 1-25 and 51-80 were written by a group of different users. Topics 26-50 were mostly written by a single user and cover the general domain of computers. Topics 81-150 were also written by a single user, but cover many domains.

Although the subject domain of the topics was diverse, some consideration was given to the documents to be searched. The initial ideas for topics were either generated spontaneously, or by seeing interesting topic areas while doing other searches. These initial ideas were then used in trial searches against a sample of the document set, and those topics that had roughly 25 to 100 hits in that sample were used as a final topic. This created a range of broader and narrower topics. After a topic idea was finalized, each topic was developed into a standardized format.

The following is one of the topics used in TIPSTER and shows the formatting.

<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 is formatted in the same standard manner to allow easier automatic construction of queries. Besides a beginning and an end marker, each topic has a number, a short title, and a one-sentence description. Then there are three major types of sections.

The first type of section is the narrative section. This section is meant to be a full description of the information need, in terms of what separates a relevant document from a non-relevant document. The narrative sections were constructed by looking at relevant documents in the trial sample and determining what kinds of information were needed to provide focus for the topic. These sections are primarily meant as instructions to the assessors, but could be used in building the queries either manually or automatically. The narratives often contain augmentations of the description, such as examples, or restrictions to focus the topic. The following three narratives illustrate these.

Example of augmentation

<num> Number: 082

A relevant document will discuss a product, e.g., drug, microorganism, vaccine, animal, plant, agricultural product, developed by genetic engineering techniques; identify an application, such as to clean up the environment or human gene therapy for a specific problem; or, present human attitudes toward genetic engineering.

Example of positive restrictions

<num> Number: 121

A relevant document will provide obituary information on a prominent U.S. person who died of an identified type of cancer. In addition to the individual's name and cancer, the report must provide sufficient biographical information for a determination of why the life and contributions of the individual were worthy of some comment upon death. In other words, a one or two line obituary is NOT sufficient.

Example of negative restrictions

<num> Number: 067

A relevant document will report the location of the disturbance, the identity of the group causing the disturbance, the nature of the disturbance, the identity of the group suppressing the disturbance and the political goals of the protesters. It should NOT be about economically-motivated civil disturbances and NOT be about a civil disturbance directed against a second country.

Many narratives are a mix of augmentation and restriction. Whereas the narratives did provide the type of clear direction needed by the human assessors, they also provide a challenge for query construction by machines.

The second type of section is the concepts section. This section is meant to reflect the "world-knowledge" brought to the task by the users, and is the type of information that could be elicited by prompts from a good interface. The concepts sections were constructed by locating "useful" information in some of the relevant documents in the trial sample. This information was then grouped into conceptually related ideas, although these relationships vary widely across topics. To show some examples of concepts, the concept lists for the three narratives previously shown are given.

<num> Number: 067
<con> Concept(s):

1. protest, unrest, demonstration, march, riot, clash, uprising, rally, boycott, sit-in
2. students, agitators, dissidents
3. police, riot police, troops, army, National Guard, government forces
4. NOT economically-motivated

<num> Number: 082
<con> Concept(s):

1. genetic engineering, molecular manipulation
2. biotechnology
3. genetically engineered product: plant, animal, drug, microorganism, vaccine, agricultural product
4. cure a disease, clean up the environment, increase agricultural productivity

<num> Number: 121
<con> Concept(s):

1. cancer
2. death, obituary

It should be noted that the number of concepts given, the organization of the concepts, and the "usefulness" of the concepts vary widely across the topics. The concepts in general provide excellent keywords for retrieval systems, although this also varies widely across the topics.

The third type of section is the optional factors section and optional definition section. These sections only appear when necessary; the factors section is an attempt to codify some of the text in the narrative for easier use by automatic query construction algorithms, and the definition section has one or two of the definitions critical to a human understanding of the topic. Two particular factors were used in the TIPSTER topics: a time factor (current, before a given date, etc.) and a geographic factor (either involving only certain countries or excluding certain countries). The definitions section was minimally used in the TIPSTER topics, but did provide some critical definitions for some of the more unusual terminology.

1.4 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 TIPSTER 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 over a million documents for each topic, resulting in over 100 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 building the TIPSTER collection, was to make relevance judgments on the sample of documents selected by the various participating systems (including the systems in TREC). 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 [2].

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 X 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

The aim in a pooling method is to have a broad enough sample of search systems so that most relevant documents for a given topic are found. Since it is well known that different systems retrieve different sets of relevant documents [3], it is critical to have as many systems as possible contributing to this pool. For that reason, the TIPSTER systems were judged against the pool of relevant documents constructed from both the TREC and TIPSTER evaluations. For the 12-month evaluation, no TREC results were available and therefore only a partial TIPSTER evaluation was possible. However the relevant documents found at the 12-month evaluation were pooled with the TREC-1 results at the first TREC conference.

At each evaluation period, the merged list of results was then shown to the human assessors. For the TIPSTER 12-month evaluation, the top 200 documents for each run were judged, but the overwhelming number of unique documents for TREC-1 meant that only the top 100 documents for each run were judged. This still resulted in an average of 1462.24 documents judged for each topic, 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. The two histograms on the next page show the number of documents judged relevant for each of the topics. The topics are sorted by the number of relevant documents to better show their range and median. The first histogram shows the number of relevant documents for topics 51-100, first used as adhoc topics against disks 1 and 2, and then used as routing topics against disk 3. The second histogram shows the number of relevant documents for topics 101-150 used as adhoc topics against disks 1 and 2.

Several interesting facts can be seen by looking at the first histogram. The median number of relevant documents per topic for disks 1 and 2 is 277, with 22 topics having 300 or more relevant documents, and 11 topics having more than 500 relevant documents. When these same topics were used as routing topics against different data (disk 3), only about half the number of relevant documents were found. This is reasonable given that there is only about half the amount of data. Some topics have far fewer than half the number of relevant documents; in general these are the topics that are particularly time dependent.

The second histogram shows the number of relevant documents for topics 101-150 against disks 1 and 2. The topics are narrower than topics 51-100 as the result of an effort to create fewer topics with more than 300 relevant documents. This effort was triggered by the concern that topics with more than 300 relevant documents are likely to have incomplete relevance assessments. Only 11 topics have more than 300 relevant documents, with only 2 topics having more than 500 relevant documents. The median number of relevant documents is 201 for this set of topics, down from 277 for topics 51-100.

1.5 Some Preliminary Analysis

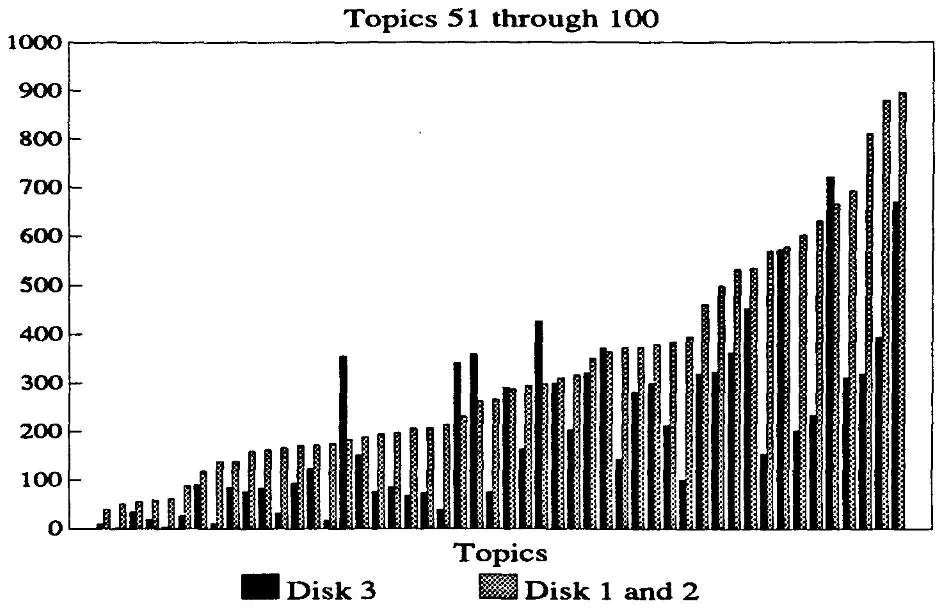
Much analysis needs to be done on this test collection. Questions about the effects of the long and varied documents, the complex topics, the pooling of system results, and the relevance judgments will be asked (and answered) as the test collection is further used and investigated. The following section discusses some preliminary answers to these questions, particularly questions relating to how well the collection has served the TIPSTER evaluation task.

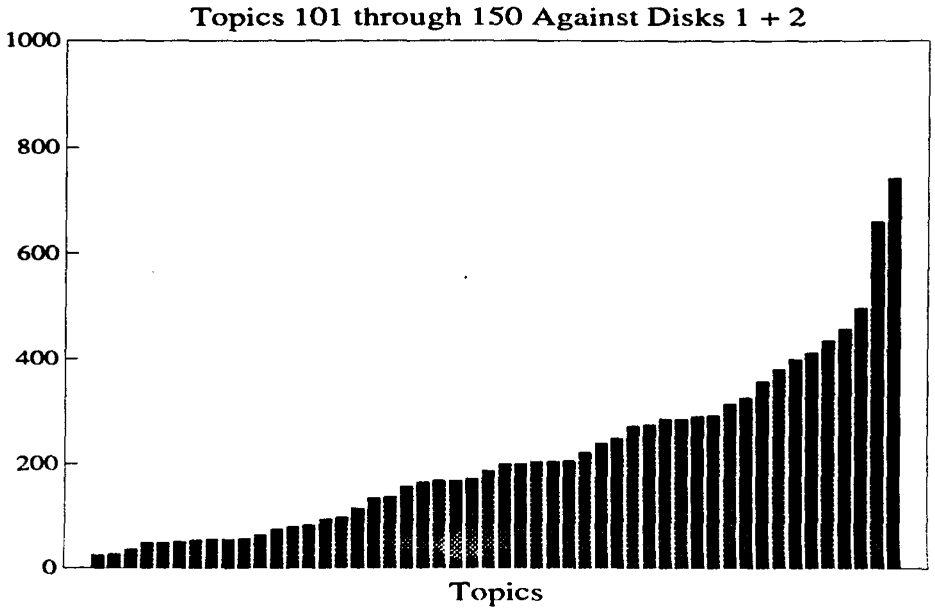
The documents

The first question involves the effect of the various sources and lengths of the documents. Table 3 shows the distribution of retrieved, judged, and relevant documents across the collections. The first 5 rows show the documents from disks 1 and 2 using topics 101-150. The last 4 rows show the documents from disk 3 using topics 51-100. The second column shows the total number of documents from each document source that was retrieved by rank 100 by any of the systems in TIPSTER and TREC-2. The third column shows how many of these documents were unique and therefore were judged. The final column shows how many of these documents were actually relevant.

Database	Retrieved	Judged	Relevant
AP	96135	16530 (17%)	4823 (29%)
DOE	15544	4018 (26%)	678 (17%)
FR	24848	11818 (48%)	410 (3%)
WSJ	125921	23706 (19%)	4556 (19%)
ZIFF	19886	6770 (34%)	1183 (17%)
AP3	187037	49304 (26%)	5848 (12%)
ZIFF3	57580	25577 (44%)	2668 (10%)
PAT	12070	4656 (39%)	146 (3%)
SJMN	161102	37695 (23%)	2322 (6%)

Table 3: Distribution of top 100 documents across databases


Topics 51 through 100
1000
900
800
700
600
500
400
300
200
100
0
Topics
Disk 3
Disk 1 and 2


Topics 101 through 150 Against Disks 1 + 2
1000
800
600
400
200
0
Topics

The analysis of the adhoc topics 101-150 (first 5 rows) shows that by far the largest number of relevant documents come from the document sources covering all domains (AP -- 4823; WSJ -- 4556). Of the five document sources on disks 1 and 2, these two had the overwhelmingly highest number of relevant documents, the lowest percentage of unique documents, and the highest percentage of relevant documents found in the judged set. The very long FR documents had few relevant documents, but a high number of retrieved documents and a high percentage of unique documents. This demonstrates the difficulty most retrieval systems had in screening out long documents and shows the almost random nature of retrieval from this set of "noise" documents. The much shorter DOE documents caused no such problems, with a larger number of relevant documents being found, but with a far fewer number of documents being retrieved. The lower percentage of unique documents for the DOE documents as opposed to the FR documents indicates that these short documents are being effectively handled by the systems. The single-domain ZIFF document source also appears to be as effectively retrieved as the all-domain sources.

The analysis of the routing topics 51-100 (last 4 rows) shows much the same pattern. The long PAT documents cause the same high retrieval and low accuracy as the FR documents did for the adhoc. Additionally, the number of unique documents is a higher percentage for all data sources in the routing task. This is due to the broad range of query terms used in routing, where the query terms are usually taken from both the topic and the training documents. The broad range of terms used across the systems caused more unique documents to be retrieved. However, a smaller percentage of these documents are relevant, likely because of time differences in the test data.

Many of the same trends can be seen if this distribution is broken down by topics. Tables 4 and 5 show the distribution of the relevant documents across databases on a per topic basis. Table 4 shows the first set of topics, used for routing in TREC-1. The first 25 topics are mostly based on information in the Wall Street Journal, with many in the financial domain. Almost all of the second 25 topics (26-50) are in the single domain of computers. As would be expected, the first 25 topics have mostly WSJ documents as relevant, whereas the second 25 topics have mostly ZIFF documents as relevant. Very few of these topics had any DOE or FR relevant documents. Addit;ionally the first 25 topics have many AP documents as relevant, as many of the financial topics are discussed in the newswire. The second 25 topics have few AP relevant documents in general, but a reasonably large number of WSJ documents since computers are often discussed in the Wall Street Journal.

Topic	AP	DOE	FR	WSJ	ZIFF
1	111	0	12	234	13
2	215	0	0	273	16
3	74	3	6	386	125
4	41	0	0	136	0
5	40	0	47	145	53
6	129	0	0	240	1
7	158	1	1	174	3
8	29	2	0	212	7
9	53	0	3	85	0
10	148	0	5	315	1
11	231	5	0	116	20
12	272	24	13	152	3
13	71	23	8	148	6
14	53	0	26	234	0
15	58	0	0	41	0
16	43	1	2	144	0
17	144	5	209	139	0
18	69	0	0	143	0
19	61	0	42	185	13
20	63	0	5	370	153
21	23	4	0	51	17
22	515	0	18	191	3
23	142	1	2	101	0
24	145	6	0	364	3
25	39	32	0	23	0
26	14	0	0	82	430
27	6	11	0	39	306
28	19	0	0	110	352
29	6	0	1	95	68
30	0	0	0	57	425
31	1	0	0	24	264
32	3	0	0	20	191
33	13	0	0	77	725
34	0	1	1	19	452
35	1	0	0	17	397
36	6	0	2	4	290
37	7	0	1	10	548
38	216	0	0	74	0
39	0	2	1	13	780
40	148	0	129	199	1
41	8	1	2	23	159
42	78	5	3	161	866
43	43	11	33	64	63
44	57	0	0	130	154
45	52	1	10	44	364
46	38	0	0	14	62
47	36	0	0	225	189
48	19	0	0	103	139
49	25	3	0	74	134
50	5	0	1	5	27

Table 4: Number of relevant documents by database

Topics 51-100 shown in table 5 were used as adhoc topics in TREC-1 and TIPSTER-12 month, and as routing topics in the later evaluations. Topics 101-150 were only used as adhoc topics. Both these sets of topics cover many domains, and have equally large numbers of relevant documents from both the AP newswire and the Wall Street Journal. They have few relevant from either DOE or FR, and those relevant are concentrated in about 11 topics out of the 100. The relevant documents from ZIFF are also concentrated in general, with slightly more topics having relevant ZIFF documents than FR or DOE documents. Note that the count of the relevant AP and ZIFF documents shown for topics 51-100 include documents from all three disks. However the WSJ documents from disks 1 and 2 reflect the use of the topics for the adhoc runs, whereas the SJMN documents are used only for routing. This may explain why many of the topics have high numbers of relevant documents from the WSJ set, with few from the SJMN set. The time difference of these sets could affect some topics, or this difference could be caused by a slight difference in the domains covered by these newspapers.

The topics

The topics are both very long and very complex, and have many more relevant documents than other test collections. A major question therefore is how these characteristics affect retrieval performance. Table 6 shows a preliminary analysis of system performance with respect to topic broadness and the level of restrictions and the use of factors in the topics. The difficulty or "hardness" measure shown in column two is the average relative recall for that topic across all systems in TREC and TIPSTER that used the full document collections. The relative recall is defined as the recall at R relevant documents (where R is the total number of relevant documents for a topic) if R is less than some threshold (100 in this case), OR the precision at that threshold if R is greater than the threshold. This measure shows how effectively a system is performing at the early stage of retrieval, and is to a large extent independent of the total number of relevant documents found for a topic. The average of this measure over many systems is a good indication of the difficulty of a given topic.

Column 3 of table 6 shows the total number of relevant documents for a given topic. A strong relationship can be seen between the difficulty of a topic and the number of relevant documents found for topics 101-150 in that the narrower topics are also the harder topics. Column 4 shows the type of restrictions used in the narrative section of a topic. The "R" stands for some type of restriction whereas the "N" means that a "NOT" was explicitly used in the topic narrative (e.g." A document is NOT relevant if it merely mentions the illegal spread of a computer virus ..."). There is a possible weak relation between the restrictions and the performance, but this trend seems to be more related to the broadness of a topic, i.e. narrow topiscs are more likely to have restrictions in the narrative. The final column contains a code for any factors used in the factors section of the topic--"G" stands for geographic factors and "T" stands for time factors. There is no clear relationship between the use of factors and the difficulty of a topic.

Table 7 shows the same statistics but for topics 51-100 as used in TREC-1 and the earlier TIPSTER evaluations as adhoc topics. Again there is no relationship between the use of factors and topic difficulty. The relationship between the use of restrictions and topic difficulty is not obvious for this set of topics, and the relationship between restrictions and topic broadness has also disappeared. Finally the relationship between topic broadness and topic difficulty that is clearly seen for topics 101-150 is very weak for topics 51-100.

What is the explanation for these observations? First, the lack of correlation between the use of factors and retrieval performance, and the weak correlation between the use of restrictions and retrieval performance reflects the idiosyncratic use of language. The performance of retrieval systems on a given topic depends on the ease with which a topic maps to the document set. If a topic contains much the same terms as do its relevant documents, then the match is easier than for a topic which has more general terminology or one that requires inferences to make the map to many of the relevant documents. For example, the description for topic 74 reads "Document will cite an instance in which the U.S. government propounds two conflicting or opposing policies". Clearly this topic will be difficult, regardless of the presence of any type of restrictions or factors. In contrast, the topic description for topic 87 is "Document will report on current criminal actions against officers of a failed U.S. financial institution". This topic not only has more specific terminology, but more readily lends itself to term expansion techniques.

The relationship of topic difficulty and topic broadness is more tentative. Whereas there is a strong relationship for topics 101-150, the relationship is weak to non-existent for topics 51-100. Three different explanations need further exploration. First, there may be an overall difference between the topic sets that remains to be characterized. Second, there may be a relationship between the maturity of retrieval systems and the correlation of topic hardness and broadness. The overall performance level used to measure topic hardness for topics 101-150 was much improved from that used for topics 51-100. And third, it

Topic	AP	DOE	FR/PAT	SJMN	WSJ	ZIFF	Topic	AP	DOE	FR	WSJ	ZIFF
51	36	0	1/0	6	104	2	101	27	17	2	7	6
52	631	2	0/0	107	251	13	102	19	33	3	7	2
53	389	0	0/0	33	296	7	103	55	0	14	25	0
54	157	0	0/0	9	63	66	104	51	0	3	25	0
55	518	0	0/0	41	564	8	105	33	0	0	26	0
56	549	0	1/0	131	591	10	106	47	0	27	146	3
57	110	0	0/0	22	177	471	107	30	0	3	65	0
58	144	0	0/0	30	60	1	108	94	0	11	189	36
59	926	0	0/0	175	72	0	109	8	1	15	219	560
60	24	0	0/0	2	45	8	110	387	0	0	150	1
61	104	0	0/0	29	139	2	111	113	56	3	124	2
62	590	0	0/0	70	70	0	112	12	11	7	334	5
63	6	10	0/0	1	6	264	113	32	6	0	72	126
64	464	1	0/0	119	88	1	114	123	4	1	20	3
65	0	68	0/4	0	20	514	115	79	0	21	85	2
66	2	33	0/0	4	3	245	116	17	0	6	28	0
67	758	0	0/0	91	92	0	117	29	0	3	106	148
68	111	96	3/2	10	49	0	118	198	1	4	87	0
69	16	3	2/0	1	31	0	119	239	3	11	84	4
70	60	0	0/0	15	15	0	120	40	0	5	48	2
71	561	0	0/0	63	72	0	121	48	0	0	2	5
72	85	0	0/0	54	74	0	122	20	5	7	86	2
73	371	0	0/0	93	77	2	123	70	156	103	106	8
74	504	9	30/0	143	273	3	124	58	31	4	77	6
75	37	64	25/12	26	86	493	125	95	0	15	76	0
76	175	2	23/0	52	191	17	126	174	11	4	57	0
77	180	0	4/0	24	19	0	127	165	1	4	68	0
78	200	2	1/0	12	28	2	128	56	7	3	296	43
79	452	5	0/0	19	98	2	129	167	0	0	49	6
80	292	1	0/0	24	205	0	130	228	0	3	64	1
81	55	0	0/0	2	9	0	131	6	0	0	22	0
82	217	84	72/35	71	317	10	132	137	0	0	64	2
83	450	137	119/0	39	125	9	133	13	10	0	29	28
84	76	175	68/22	29	130	4	134	8	140	7	23	10
85	1144	2	0/0	184	268	12	135	75	183	11	156	4
86	27	0	5/0	33	190	1	136	10	0	0	121	92
87	162	0	0/0	46	136	0	137	54	0	0	111	2
88	99	0	0/0	0	99	0	138	36	0	0	20	0
89	71	1	1/0	4	115	0	139	47	0	0	10	0
90	98	26	6/0	4	206	1	140	25	0	0	6	0
91	12	0	0/0	2	35	0	141	22	0	0	16	0
92	67	0	0/0	6	41	2	142	336	2	54	338	3
93	198	0	1/0	46	20	2	143	271	0	45	135	1
94	119	0	0/0	92	67	347	144	42	0	0	7	0
95	92	0	63/4	95	52	341	145	103	0	0	64	0
96	40	490	13/60	63	73	284	146	320	0	0	68	0
97	40	10	6/6	38	90	488	147	118	0	2	209	16
98	38	3	1/1	31	186	1157	148	220	0	1	77	0
99	392	0	0/0	80	115	0	149	30	0	1	98	49
100	107	3	64/0	51	95	207	150	236	0	7	254	5

Table 5: Number of relevant documents by database

Topic	"Hardness"	No. relevant	Restrictions	Factors
121	0.0471	55	R	GT
120	0.0893	83	N	
141	0.1552	36	N	G
139	0.1759	55	N	
140	0.1765	25	N	
101	0.1925	57	N	G
114	0.2032	138	N	
149	0.2153	135	N	
124	0.2176	173		
131	0.2269	28	N	
122	0.2341	114	R	
102	0.2362	64	N	G
138	0.2472	52		
113	0.2568	206	N	
144	0.2749	49	N	
105	0.2849	54	N	G
107	0.2911	98		G
147	0.3165	315	N	G
116	0.3271	49	N	
125	0.3285	169	N	
106	0.3479	201	N	G
117	0.3538	275	N	GT
104	0.3612	75		G
127	0.3647	223	N	G
118	0.3665	273	R	
119	0.3688	326	R	
108	0.3718	294	N	G
112	0.4044	291		
115	0.4344	165	N	
136	0.4538	206	N	
145	0.4538	169		G
137	0.4544	158	N	G
129	0.4771	207	N	
128	0.5050	381	N	T
143	0.5291	412		G
123	0.5585	435		
134	0.5924	188	N	
133	0.6162	80	N	
126	0.6297	240	N	
132	0.6359	201	N	G
130	0.6574	286		
111	0.6991	286		
109	0.7065	742		
135	0.7609	400	N	
142	0.7809	660		
146	0.7953	358		
150	0.7956	458		G
110	0.8200	496	R	
110	0.8200	94	R	
148	0.8935	250		

Table 6: Characterization of topics by "hardness", broadness, and levels of restrictions

Topic	"Hardness"	No. relevant	Restrictions	Factors	Routing "Hardness"	No. Routing Relevant
74	0.1140	499	N	G	0.4444	323
93	0.1407	171	R		0.2245	94
61	0.1811	206			0.2500	67
88	0.1833	166	R		0.3878	32
73	0.1893	183	R		0.3255	355
98	0.2280	666	R		0.2688	722
90	0.2370	266			0.1746	75
85	0.2447	896	R		0.1600	670
76	0.2493	294	R	G	0.2763	163
82	0.2615	602			0.5625	203
96	0.2807	693	R		0.4394	310
81	0.2915	62	R		0.1109	4
64	0.2940	375	R		0.3049	282
89	0.2967	175			0.6777	17
92	0.3051	88	R	T	0.2917	27
69	0.3099	52	R		0.4622	1
95	0.3160	263			0.4019	359
91	0.3194	40	R	GT	0.1608	9
75	0.3433	365			0.3069	372
67	0.3773	534	N	T	0.2209	365
72	0.3813	119	R	G	0.3919	91
80	0.4007	374	R	G	0.7147	143
63	0.4020	208	R		0.7134	74
68	0.4100	195	R		0.4747	76
87	0.4260	188	R	GT	0.4016	151
77	0.4320	139	R		0.2288	85
65	0.4367	386	R		0.5103	214
78	0.4440	162	R		0.3743	83
70	0.4481	55	R		0.3438	34
94	0.4620	310	N		0.5884	300
99	0.4736	288	R		0.6000	291
66	0.4820	197	R	G	0.2931	86
59	0.4947	579	R		0.5263	574
58	0.5220	159	R	T	0.6584	76
55	0.5307	810	R		0.5438	320
97	0.5333	352	R		0.4725	319
52	0.5487	535	R	G	0.8580	454
54	0.5487	171			0.4816	124
62	0.5507	298	N	T	0.3647	426
60	0.5653	60	R		0.7328	18
100	0.5877	316			0.8984	204
51	0.5947	138			0.7716	11
79	0.5980	232	0	G	0.6261	341
84	0.6067	396	R	GT	0.3203	101
86	0.6480	214	R	GT	0.7016	40
71	0.6544	380	N		0.8134	300
83	0.8133	633	R		0.6441	235
56	0.8393	878			0.7872	395
53	0.8547	571	R	O	0.9366	154
57	0.9347	461	N	T	0.7581	319

Table 7: Characterization of topics by hardness, broadness, and levels of restrictions, both as adhoc topics and as routing topics

may be that the hard topics have less complete relevance judgments than the easy ones (however this does not explain the difference between the topic sets). All three of these hypotheses are currently under investigation.

The last section of table 7 shows the performance of topics 51-100 as routing topics. Column 6 shows the relative recall for each topic for routing against new data (disk 3). There is a weak correlation between topic difficulty as an adhoc topic, and its difficulty as a routing topic. In particular, topics that are easy as adhoc topics are still generally easy as routing topics. However some topics that were hard adhoc topics became easy routing topics, usually because of the term expansion available using the training documents (the relevant documents from the adhoc task). Some hard topics remained hard, and some topics became more difficult because of time differentials between the training data and the test data. There is a minimal correlation between the number of relevant documents as adhoc topics (i.e. the number of documents available for training) and the performance of the topic as a routing topic.

The TIPSTER topics are not only broad and complex, but also very long. Other test collections have much shorter topics of one or two sentences in length, whereas the TIPSTER topics are generally about a page in length. Longer topics provide more information, but also a large number of non-discriminating terms. One section of the TIPSTER topics, the description section, could be viewed as a one-sentence summary of the topics, and table 8 shows the difference in performance for one of the TIPSTER systems using just the description section vs. the "full" topic (results from the TIPSTER 24-month INQUERY system, INQ009 using the title, description and concepts and INQ013 using the description only). Use of the description only cuts retrieval effectiveness by about one half, both on average across the entire recall range and at the high performance end (precision at 100). The recall performance is particular affected, as would be expected. This drop in performance is due to fewer topic terms being available for matching. Interestingly, however, the shorter topics do retrieve some relevant documents not retrieved by the longer ones, although this is clearly a ranking effect since the terms from the description are a subset of the full topic terms.

It should be noted, however, that this performance difference is not due solely to topic length. The concepts section of the topic consistently provides very valuable terms for retrieval, and minimal additional improvement comes from adding the longer narrative section. This may be caused by a temporary lag in research in how to properly handle this narrative section.

	"Full" topic	Description only
Average precision	0.3465	0.1420
Precision at 100	0.4934	0.2730
Total relevant Retrieved	8688	4953
Unique relevant (total)	3794	59
Unique relevant at 100	1651	549

Table 8: Performance differences using the full topics vs the description only

The relevance judgments

Two particular issues are being investigated with respect to the relevance judgments. The first is the effectiveness of the merging of results across the TIPSTER and TREC systems to form the pool of documents sent to the relevance assessors. One of the measures of this effectiveness is the overlap of retrieved documents. If there is a heavy overlap in that all systems retrieve the same documents for a given topic, then it is likely that relevant documents will be missed.

Table 9 shows the overlap statistics for TREC-1, with a total of 28 systems including the TIPSTER systems.

	Top 200		Top 100	
	Pos.	Act.	Pos.	Act.
Unique Documents Per Topic Adhoc, 45 runs 18 groups	9000	2398.4	4500	1278.86
Unique Documents Per Topic Routing, 26 runs 17 groups	5200	1932.42	2600	1066.86

Table 9: Overlap of submitted results (TIPSTER 12-month + TREC-1)

The first overlap statistics are for the adhoc topics and the second statistics are for the routing topics. For example, out of a maximum possible 9000 unique documents (45 runs times 200 documents), about 27% of the documents were actually unique for the adhoc topics. There are significantly more unique documents for routing. The overlap is much less than was expected, which means that the systems were finding different documents as likely relevant documents for a topic. Whereas this might be expected 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. More importantly, however, many systems used different sets of keywords in the constructed queries, resulting in the often unique sets of retrieved documents. This is particularly true for the routing topics, where often unique terms were added from the training documents. The fact that the routing topics are run against only half the number of documents as the adhoc topics, but still have many more unique retrieved documents per topic suggests strongly that the wide variation in query terms is the cause of the small overlap. The same lack of overlap still holds at the 100 document mark, indicating that the systems retrieve different documents even at the beginning of the ranked lists.

This lack of overlap improves the coverage of the relevant set, and verifies the use of the pooling methodology to produce the sample. It should be noted, however, that because complete judgments were not made, the recall measures must be considered to be relative rather than absolute. Whereas each system is fairly evaluated against this pool, future systems (with widely differing methods) might find additional relevant documents that would be considered non-relevant by default.

	TREC-2		TIPSTER 24-mo.	
	Pos.	Act.	Pos.	Act.
Unique Documents Per Topic Adhoc, 52 runs 24 groups	5200	1106.0	1900	144.7
Unique Documents Per Topic Routing, 52 runs 25 groups	5200	1465.6	3600	288.8

Table 10: Overlap of submitted results (TIPSTER 18-month + TREC-2) plus TIPSTER 24-month

Table 10 shows the overlap of the submitted results from the later TIPSTER evaluations and the TREC-2 evaluation. In this case only the top 100 documents are shown as these were the only ones judged due to limited time. Note that there is a significantly higher overlap in the TREC-2 results (including the TIPSTER 18-month runs). Only about 21% of the retrieved documents for the adhoc topics were actually unique, with slightly more unique documents for the routing topics. This is likely due to better performance at this later date (fewer errors with less spread of retrieved documents). In particular, the fewer unique documents for the routing topics is probably due to the better training data in addition to better systems.

The many additional runs made for the TIPSTER 24-month evaluation on the same data as for TREC-2 resulted in few new documents. For example, out of a possible 3600 unique documents retrieved for routing (per topic), only about 8% had not already been seen in TREC-2. This implies that, at least for the current systems in TIPSTER and TREC, some asymptote is being reached for finding new documents.

This hypothesis will be investigated in the coming months. A series of experiments is being planned to investigate the coverage of the relevant documents, i.e. how many relevant documents have been missed because they did not appear in the pool for judgment. Two types of experiments are planned involving additional relevance judgments to be made for some topics. The topics will be selected to cover a range of narrow and broad topics, with an emphasis on the mid-range topics which constitute the majority of the TIPSTER topics (mid-range being around 200 relevant documents). First, systems that seem likely to retrieve unique relevant documents beyond the current 100 mark will be selected and a new pool of possible relevant documents will be created for the subset of topics being investigated. This new pool will be submitted for assessment to the same assessor used for the earlier judgments. Additionally topics will be submitted to human intermediaries to conduct extensive searches. The results from these searches will also be assessed by the official relevance assessors. These two sets of experiments will provide some indication of the completeness of the pool.

An issue related to the completeness problem is the issue of relevance bias. Whereas it can be assumed that the judgments are not absolutely complete, it is hoped that the judgments are not biased toward one type of system as opposed to another, e.g. a system using a particular type of weighting scheme. This is a possibility given that the TIPSTER contractors contributed many more documents for assessment than the TREC participants (and therefore the assessments may be more biased toward the TIPSTER systems), or alternatively many of the TREC systems were based on a single weighting scheme, and this could cause a bias. The existence of this bias will also be investigated.

In addition to the completeness or bias of the pooling method, it is necessary to investigate the quality of the relevance assessments themselves. Two different consistency issues must be addressed-- the consistency of a single judge and the consistency between judges on a single topic. The first issue will be addressed by submitting random

Subset of collection	WSJ	AP	ZIFF	FR	CRAN
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
Number of unique terms	156,298	197,608	147,405	116,586	8226

Table 11: Comparison of the TIPSTER collection to other collections

samples of judged documents to the same judge for new assessments and the second issue will be addressed by submitting the same random samples to a different judge for assessments.

1.6 Comparison to Other Collections

The TIPSTER collection is many magnitudes of scale larger than existing collections in several ways. First there are many more documents. Table 11 shows a comparison of the documents on disk 1 of the TIPSTER collection (about one-third of the full collection) to the Cranfield collection, ne of the oldest collections. Even the largest publically-available collection, the NPL collection, only has 11,429 documents. Second, the documents are longer, with most documents being full text as opposed to the abstracts found in other collections.

The topics are also longer and more complex. This issue was discussed earlier in the analysis section, and therefore will not be repeated here. However the long and complex topics are the main research challenge of the TIPSTER collection; the difference in document collection size and document length generally present system engineering challenges and cause less difficulty after the initial system scaling-up.

There are also many more relevant documents per topic. Older collections tend to have an average of ten or fewer relevant documents per topic, whereas the average number of relevant documents per topic for the TIPSTER collection is over 200.

2. THE JAPANESE TEST COLLECTION

The design of the Japanese test collection was similar to that of the English collection, with similar formats used in both the documents and topics. However because of the difficulty in obtaining large amounts of Japanese data, the Japanese part of the project was slower to start. The documents, 206 megabytes (151,650 records), of the Nikkei newspaper were distributed in the second year of the project. Topic development was also delayed due to difficulty in discovering methods of constructing topics. Seven topics were delivered in time for the 24-month evaluation, and results from the two TIPSTER contractors working in Japanese were assessed for relevance. The development of a full collection in Japanese is a continuing effort.

3. REFERENCES

[1] Belkin N.J. and Croft W.B. Retrieval Techniques. In Williams, M. (Ed.), *Annual Review of Information Science and Technology* (pp. 109-145). New York, NY: Elsevier Science Publishers, 1987.

[2] 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.

[3] 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.