INQUERY Does Battle With TREC-6

James Allan, Jamie Callan, W. Bruce Croft,
Lisa Ballesteros, Don Byrd, Russell Swan, Jinxi Xu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts USA

This year the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts participated in eight of the ten tracks that were part of the TREC-6 workshop. We started with the two required tracks, ad-hoc and routing, but then included VLC, Filtering, Chinese, Cross-language, SDR, and Interactive. We omitted NLP and High Precision for want of time and energy.

With so many tracks involved, it is nearly inevitable that something will go wrong. Despite our best efforts at verifying all aspects of each track—before, during, and after the experiments—we once again made mistakes that were minor in scope, but major in consequence. Those mistakes affected our results in Ad-hoc and Routing, as well as the dependent tracks of VLC and Filtering. The details of the mistakes are presented in each track's discussion, along with information comparing the submitted runs to the corrected runs. Unfortunately, those corrected runs are not included in TREC-6 summary information.

This remainder of this report covers our approach to each of the tracks as well as some experimental results and analysis. We start with an overview of the major tools that were used across all tracks. The paper is divided into the following sections. The track descriptions are generally broken into approach, results, and analysis sections, though some tracks require a different description.

1. Tools applied (Inquery, InRoute, LCA)
2. Ad-hoc track
3. Routing track
4. Very Large Corpus (VLC) track
5. Filtering track
6. Chinese track
7. Cross-language IR (CLIR) track
8. Spoken document retrieval (SDR) track
9. Interactive track

A. CLIR track questionnaire
B. TREC interactive track protocol log

1 Tools applied

Although UMass used a wide range of tools, from Unix shell scripts, to PC spreadsheets, three major tools were applied across almost all tracks: the Inquery search engine, the InRoute filtering engine, and a a query expansion technique known as LCA. This section provides a brief overview of each of those so that the discussion does not have to repeated for each track.

1.1 Inquery

All tracks other than the filtering track used Inquery[9] as the search engine, sometimes for training, and always for generating the final ranked lists for the test. We used Inquery V3.1 or V3.2. The former is the most recent version of Inquery made available by the CIIR; the latter is an in-house development version. The differences between the two are not consequential for this study.

The current belief function used by Inquery to calculate the belief in term t within document d is:

$$w_{t,d} = 0.4 + 0.6 \times \frac{\mathrm{tf}_{t,d}}{\mathrm{tf}_{t,d} + 0.5 + 1.5\frac{\mathrm{length}(d)}{\mathrm{avg\ len}}} \times \frac{\log \frac{N+0.5}{n_t}}{\log N + 1}$$

where n_t is the number of documents containing term t, N is the number of documents in the collection, "avg len" is the average length (in words) of documents in the collection, length(d) is the length (in words) of document d, and $\mathrm{tf}_{t,d}$ is the number of times term t occurs in document d.

1.2 InRoute

InRoute is a varient of Inquery modified to be more efficient for processing large numbers of queries on a stream of documents [8]. As a filtering engine, it processes the incoming documents one at a time. It does not have access to statistics about the incoming collection, but can use a retrospective collection for any statistics needed. InRoute has the ability to learn collection statistics as documents stream by, and can also use relevence judgements to refine a query incrementally as the training documents arrive.

Inroute was used only in the filtering track.

1.3 Local Context Analysis (LCA)

In SIGIR '96, the CIIR presented a new query expansion technique that worked more reliably than previous "pseudo relevance feedback" methods.[13] That technique, Local Context Analysis (LCA), locates expansion terms in top-ranked passages, uses phrases as well as terms for expansion features, and weights the features in a way intended to boost the expected value of features that regularly occur near the query terms.

LCA has several parameters that affect its results. The first is the choice of LCA database: the collection from which the top ranked passages are extracted. This database could be the test collection itself, but is often another (perhaps larger) collection that it is hoped will broaden the set of likely expansion terms. In the discussion below, if the LCA database is not the test collection itself, we identify what collection was used.

LCA's other two parameters are the number of top passages used for expansion, and the number of expansion features added to the query. In all cases, the LCA features were put into a query construct that allows a weighted average of the features. Assuming n features, f_1 through f_n, they are combined as:

$$\begin{array}{lcl} \texttt{\#wsum(1.0} & 1.0 & f_1 \\ & \vdots & \vdots \\ & 1-(i-1)*0.9/s & f_i \\ & \vdots & \vdots \\ & 1-(n-1)0.9/s & f_n\ \texttt{)} \end{array}$$

Here, s is scaling factor that is usually equal to n. The weighted average of expansion features is combined with the original query as follows:

$$\texttt{\#wsum(1.0}\ 1.0\ \text{original-query}\ w_{\text{lca}}\ \text{lca-wsum)}$$

where w_{lca} is the weight that the LCA features are given compared to the original query. Note that the final query is a weighted combination of the original query and the expansion features.

2 Ad-hoc track

The focus of the research carried out for the adhoc track was on query processing, query expansion, weighting and core concept identification. Most of this work was expected to produce incremental improvements compared to the techniques used in previous years, although the core concept research continues a new direction in the use of the Bayesian net model.

The official results in the ad-hoc track are significantly lower than they should be because of a failure to index Volume 5 of the test data.

2.1 Ad-hoc approach

In the query processing area, the emphasis was to produce a simpler, but effective process to replace the rather complex mixture of linguistic and statistical techniques that had been developed for TREC in previous years. The three steps in the new process are removing "stop structure", identifying phrases and proper nouns, and recognizing the presence of foreign country requirements. Stop structure refers to language constructs that are often found in queries as "fillers" and which can have occasional negative effects on retrieval. Examples of such structure are "give me documents on...", "pros and cons of...", "a relevant document will contain...", and "I am interested in...". Stop structure removal uses a table of such structures, and this part of query processing was only a minor modification of the previous year's process.

Phrase identification this year was based primarily on a phrase dictionary, rather than the part of speech tagging that was used previously. To construct this table, a lexical acquisition program was created to process large amounts of text and select suitable phrase candidates. Both part of speech and statistical approaches to identifying phrases were used, but our evaluations shows that the statistical approach was both faster and more accurate. The statistical approach, which is very similar to the statistical phrases first used by Salton in the 1970s, records phrase candidates, refines them, and then removes those with low frequencies. The phrase candidates are sequences of non-stop words, where stop words include the usual small list of words used in many retrieval systems plus irregular verbs, numbers (with some exceptions), dates, some punctuations, title words, company designators and locations. Long sequences of words are then split using rules that look for certain endings, case changes, conjunctions, and hyphenations. A final refinement checks to see if subsequences can replace longer sequences. The phrase table is then used at query processing time to identify all possible phrases in the query. Phrases are represented using the INQUERY model which decides how significant the proximity component of the phrase is and also looks for phrase words to occur in passages. This is represented as `#passage25 ( #phrase(` *words* `))`.

For query expansion this year, we investigated refinements of the Local Context Analysis (LCA) approach first used in TREC last year and described in a recent SIGIR paper.[13] In particular, we have used different parameters for number of text passages used in the expansion and the number of concepts added to the query. In TREC-5, we found that using fewer passages (the top 20) for expansion produced better results. This was not something we observed with any other combination of database or queries. In fact, the expansion results in other tests were consistent with many more passages and 100 were used as a default in TREC-5. Although the TREC-5 queries may be unusual, we decided to be more conservative and use 30 passages this year. We also reduced the number of expansion concepts from 70 to 50. The value of w_{lca} was 1.25, meaning that the expansion features were given 125% the weight of the constructed query.

A more significant change in the LCA approach used this year was to base the expansion on passages retrieved from a larger database than just volumes 4 and 5—we used TREC volumes 1 through 5, with the Federal Register data omitted. The reason for this is simple: increasing the size of the database increases the likelihood that topical material will be retrieved and therefore increases the likelihood of finding good expansion concepts. There are two ways that this approach could negatively affect results. One is that many documents with content of little interest but containing a number of query terms could be introduced by using the larger database. Federal Register documents are a good example of such documents. In these experiments, we excluded Federal Register documents from the archive used for expansion. The other way in which a larger database could lower effectiveness is by producing documents that, although on the correct general topic, are from the wrong time period. An example would be looking for recent documents about cooperation between Iran and Iraq, but basing the expansion on documents describing the various Iran-Iraq conflicts in the last decade. This is a problem even if just volumes 4 and 5 were used, since some of the TREC queries refer to events that are more recent than any of the data. For this reason, we did not try to correct this problem by, for example, using only documents with recent dates in the expansion.

In the weighting and core concept area, we investigated a combination of weighting and clustering techniques to identify the most important concepts in a query, including both the original concepts and expansion concepts. The process used was to weight the original query words and phrases using a combination of idf and the average term frequency in the collection. This weighting method appears to give quite reliable rankings of the importance of the concept. The weight itself, however, does not produce effectiveness improvements. Instead, we simply gave the highest ranking word or phrase a higher weight (1.5) than the rest of the query. If a single word was at the top rank, we also assigned any phrase that contained the word the same higher weight. This was intended to give the core word more context from the query. One other weighting heuristic used was that if our recognizer identified the presence of a foreign country reference in the query (`#foreigncountry`), this term was assigned the higher weight. We did this to reflect the importance of these references in many of the TREC queries.

We also looked at changing the weighting of query and expansion concepts based on how they clustered. The clustering can be based on how concepts co-occur in the collection or on how they co-occur in the retrieved documents. Although this technique shows some promise, we were not able to identify a consistently reliable implementation in time for the TREC runs. We continue to look at this issue and are also looking at using more sophisticated INQUERY operators[11] to capture models of core concepts.

2.2 Ad-hoc results

Our TREC-6 ad-hoc submissions were both flawed in that they were run against only TREC Volume 4 and not Volume 5. The following discusses the results of the *corrected* runs, not the official runs. For comparison, we include the flawed runs in Table 1 and Figure 1.

The CIIR's ad-hoc query processing included three major steps:

1. Basic query processing—removing stop phrases and stop word from the description field (for INQ401) or the title and description fields (for INQ402).

2. Phrase identification.

3. Adding up to 50 features via query expansion with LCA.

For this analysis, we applied those steps to three queries: the title, the description, and a combination of the title and description (no phrase identification was done to the title-only run). Table 1 shows evaluation numbers for the nine combinations. In all cases, each successive stage of processing improves the quality of retrieval. The very short title queries out-performed the description queries almost uniformly, but their combination provided even better retrieval quality. Figure 1 shows a recall/precision graph of the three runs (the runs represented in the bottom row of Table 1).

For comparison, the average precision for the submitted INQ401 was 0.1440, a 38% drop in effectiveness because of omitting half the collection. For INQ402's submitted run the average precision was 0.1612, a 40% drop. TREC volume 4 contains 293,710 documents, compared to the 556,077 in volumes 4 and 5, so we accidentally omitted 47% of the test collection. Of the 4611 relevant documents possible for the ad-hoc track, 58% of them came from volume 5. It is intriguing that losing 47% of the collection and 58% of the relevant documents did not cause an entirely proportional drop in effectiveness.

2.3 Ad-hoc analysis

The evaluation of the ad-hoc process by component steps as illustrated in Table 1 shows that each of the components provided some value. The identification of phrases showed a modest improvement of 4-6%,

	Title	Desc (INQ401)	Title&Desc (INQ402)	Flawed INQ401	Flawed INQ402
Basic	0.2054	0.1663	0.2103		
@20	0.3320	0.2910	0.3620		
R-prec	0.2474	0.2140	0.2461		
+ phrases	0.2149	0.1937	0.2441		
@20	0.3300	0.3240	0.3790		
R-prec	0.2668	0.2345	0.2822		
+ LCA	0.2477	0.2327	0.2730	0.1446	0.1612
@20	0.3710	0.3850	0.4200	0.2620	
R-prec	0.2910	0.2817	0.3021	0.1839	

Table 1: Comparison of three phases of ad-hoc query processing on three types of starting queries. Each cell contains the average precision, the precision at 20 documents retrieved, and the R-precision, in that order from top to bottom. The last two columns contain information about the official (flawed) runs.

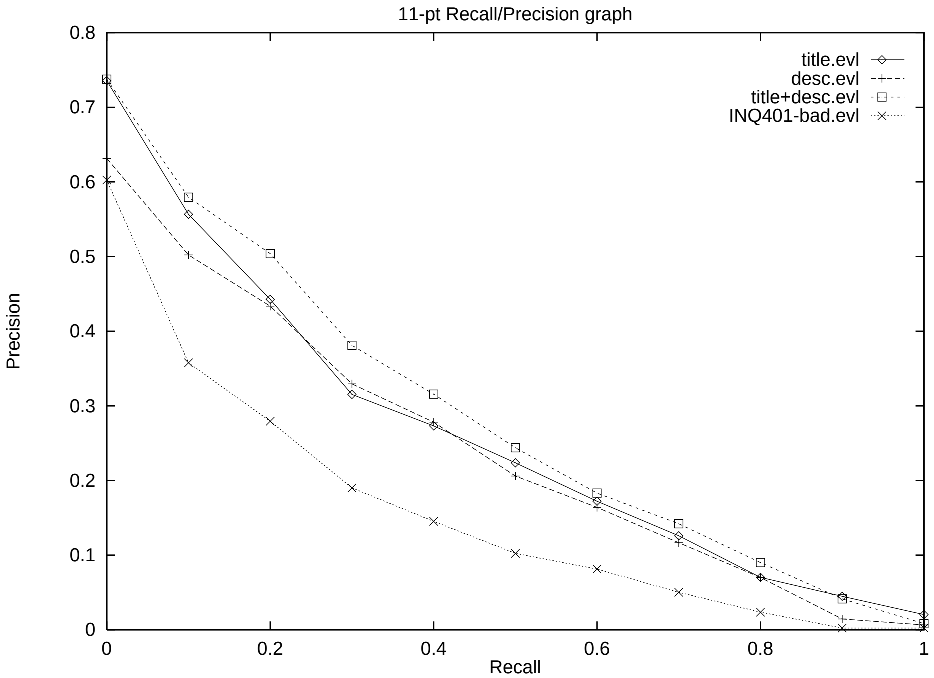

Figure 1: Recall/precision tradeoff for ad-hoc process applied to titles, descriptions, and the combination. The last two are official runs INQ401 and INQ402, respectively. (These submitted but flawed INQ401 results are provided for comparison.)

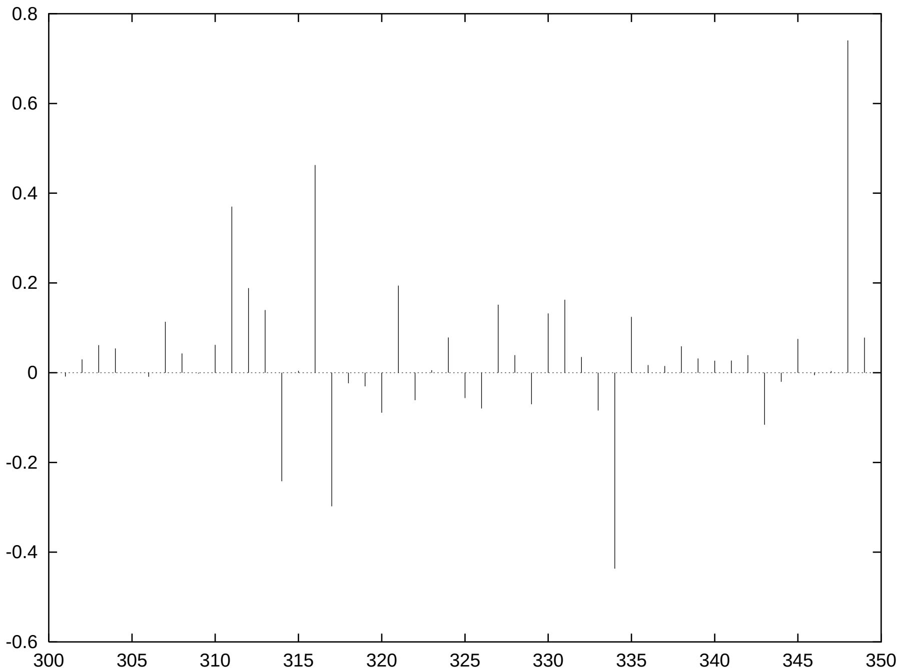

Figure 2: Change in average precision for the ad-hoc queries when the title is used as a basis for the query rather than the description. These results are for basic query processing.

though it is not statistically significant (by a sign test). The additional 15% or so improvement caused by the LCA expansion is, however, statistically significant at virtually all levels of recall and all document cutoff values.

One of the more interesting characteristics of the queries is the noticeably better effectiveness that the short, 2 or 3 word title queries achieve as compared to the longer descriptions. The difference is almost entirely wiped out by our query processing, but it remains even then. A sign test shows that the difference is statistically significant, with a P-value of 0.0325, but it is only the average precision that is significant: the difference is not significant at any standard recall point other than 0.0, nor at document cutoffs of 5, 10,15, 20, 30,100, 200, 500, or 1000.

Some quick scanning of the results shows that although most of the title queries are substantially better, there are some that are not. Figure 2 shows the difference in average precision for the queries when the titles are used rather than the descriptions (a positive number means the title query is better). The startling quality of the very short queries is not particularly surprising considering the following:

- Topic 349 is about metabolism (it showed the greatest change by using titles). The title query is "metabolism". The description provides a definition of metabolism without using the word.
- Topic 316 is about polygamy. The title is very specific. The description includes noise words that will confuse most query engines: roots, prevalence, world, today.
- Topic 311 is about industrial espionage, but the the description mentions neither industry or espionage.
- Topic 312 is about hydroponics, but the description does not mention hydroponics.

We have not investigated whether the odd query construction in fact caused any of the mistakes in the system (perhaps articles about hydroponics only occasionally mention "hydroponics"), but it seems to be the root issue in many cases.

The LCA query expansion appears to have helped in most of those cases: Topic 311 is expanded to include "espionage", 312 gains "hydroponics", 316 now includes so many references to polygamy that the noise words are lost. Topic 349 is not helped by expansion, perhaps because it fails to acquire the word "metabolism."

3 Routing track

UMass had very little research interest in the routing track this year, and unfortunately that appears to have shown in the results: a careless error in the query running caused a large number of query terms to be entirely ignored. The approach for query formulation was very similar to that taken in TREC-5, with some minor exceptions.

3.1 Routing approach

The basic approach to the routing task was similar to last year's method. The query is exanded by extracting features that occur often in the relevant documents and rarely in the non-relevant document. Feature weights are assigned as a Rocchio combination of weights in the relevant and non-relevant documents. The final weights are adjusted using Dynamic Feedback Optimization.[6]. The peculiarities of this year's approach are as follows:

- A starting query Q_0 was created from the *all* parts of the routing topic using the methods described in the ad-hoc track.
- In TREC-5, we built 8 different training databases for the 50 routing queries. Those databases represented all possible combinations of the TREC volumes on which a routing query had been evaluated in the past. The result was that when a query was run against its training database, any unjudged documents are highly likely to be non-relevant, since that database had been at least partially judged.

  For TREC-6, we made an effort to reduce that work substantially. We built one extremely large database that included TREC volumes 1 through 4, as well as the TREC-4 and TREC-5 routing volumes (there is some overlap in those volumes; documents were not indexed twice). The training documents were selected by running Q_0 against the training database and then removing any documents that were not explicitly judged (i.e., were not in the TREC relevance judgements list), resulting in the training set S_0. A second run of Q_0 retrieved the top-ranked 200-word passages in the training collection, similarly restricted to passages from judged documents, yielding P_0.
- The documents in training set S_0 were examined and all terms that were not stop words were extracted. In addition, any phrases that occurred in the set of phrases used for ad-hoc query construction were also extracted. The result was a list of words and statistically common phrases occurring in the training documents. The training passages in P_0 were also examined for all pairs of words that occurred within a window of 20 of each other inside the passages.

  The words and phrases were sorted by the proportion of relevant training documents containing the feature minus the proportion of non-relevant training documents containing it. A feature that occurred in all of the relevant documents and no non-relevant documents would have a weight of 1.0; a feature that occurred evenly in both sets would have a weight of 0.0; and so on. The 20-window words were similarly ranked.
- A query was constructed from the features of the original query, the 20 most highly weighted terms, the 20 most highly weightedphrases, and the 20 most highly weighted 20-window pairs, for a total of up to 60 features added. In no case was a feature added if its weight from above was below 0.045.

  The features were all assigned the weight:

  $$w_q + 4w_r - \frac{1}{2}w_{nr}$$

  where w_q was the weight in the original query (zero if the feature was not in the query), w_r was the average tf value of the feature in the relevant documents (*not* the average belief), and w_{nr} was the

	INQ403 (correct)	INQ403 (submitted)
Avg prec	0.3180	0.2290
Prec @ 20	0.5106	0.4617
R-prec	0.3576	0.2898

Table 2: Routing results, showing both a correct run as well as the results from the submitted run that had large amounts of the query ignored.

average tf value of the feature in the non-relevant documents (zero if the feature did not occur in the non-relevant documents).

This created query Q_1.

- Query Q_1 was run against the training collection again and all judged documents in the top 20,000 retrieved documents were used as the basis for DFO adjustment of the weights. DFO was applied in three passes, allowing the weights to increase by 100%, by 50%, and by 25%, respectively. The resulting query is Q_2.
- Q_2 was the final query submitted to NIST and run against the test collection.

The differences between TREC-5 and TREC-6 are that an *a priori* set of statistical phrases was used rather than mining the training set for common pairs of adjacent words, for pairs within a window of 5, and for pairs within a window of 50. Further, in TREC-5 the queries were expanded with up to 250 features whereas for TREC-6 we allowed only up to 60 additions.

3.2 Routing results

Unfortunately, the process of gathering retrospective statistics for various idf values of features contained a bug. The result was that large numbers of query features were treated as if they did not occur in the database—e.g., for topic 1, 46 of 118 features were dropped from the query, resulting in a 25% drop in average precision (for that topic).

Table 2 and Figure 3 show the results of the routing run. In both cases, the submitted run is included along with the corrected run for comparison. (The 25-40% improvement from the bad run to the good run is statistically significant at all levels after the top 15 documents are retrieved.)

3.3 Routing analysis

Beyond error analysis to determine why the results were so bad, no work has been done at this time to understand how the routing query formulation worked.

4 Very Large Corpus (VLC) track

Our goal for the Very Large Corpus (VLC) track was to build and search a single database of 20 gigabytes (GB). Inquery had been tested elsewhere on databases of comparable size, so we did not expect size to be a problem. We were interested primarily in studying the times required to index and retrieve documents from a 20 GB database.

4.1 VLC approach

The indices were built in two stages. In the first stage, during document parsing, a series of temporary files were written that each contained one or more blocks. Each block was a set of inverted list fragments. When all document files had been parsed, the second stage began. In the second stage, temporary files were merged, yielding a final inverted index.

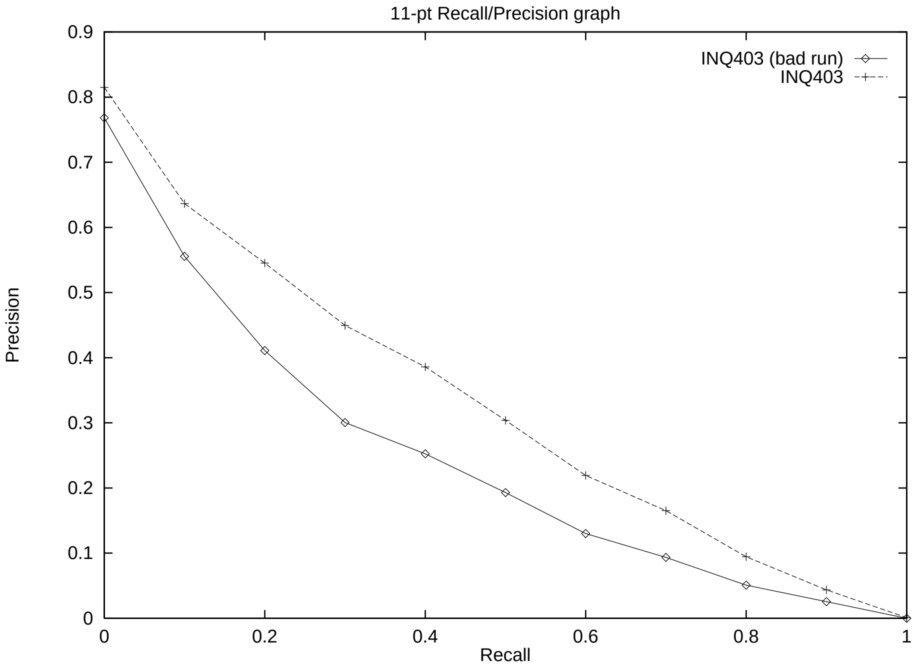

Figure 3: Recall/precision graph for INQ403, the routing run. Both the submitted and correct runs are shown for comparison.

The times required to build the 2 GB Baseline and the 20 GB Full VLC database are shown below. The figures do not include the time required to copy files from CD-ROM or DAT tape, nor the time required to uncompress the files. The experiments were run on an UltraSparc computer with 4 processors and 1 GB of memory, primarily because that machine had the (ample) disk space required for indexing the VLC corpus. Only one processor and less than 100 MB of memory were used.

Task	2 GB Time (hh:mm:ss)	20 GB Time (hh:mm:ss)	% CPU
Parse	5:55:29	61:21:16	97%
Merge	36:51	4:40:53	71%
Total	6:32:20	66:02:09	

The 2 GB index was built at a rate of 308 MB per hour, while the 20 GB index was built at a rate of 303 MB per hour. It is encouraging that indexing time scaled linearly. However 300 MB per hour is slower than expected, so we view these figures with caution.

Instead of creating new queries for the VLC track, we used the queries created for the ad-hoc track (see Section 2).

4.2 VLC results

Timing and accuracy figures are shown below for two official and four unofficial runs. The timing figures were obtained after "warming-up" the system by running query 251 from the INQ301 query set used in TREC-5. Each query returned 20 documents, as specified in VLC track guidelines.

		Full Index			Top-Docs, 1K		Top-Docs-Only, 1K	
Database	Query Set	Run ID	Time per qry (m:ss)	Prec 20	Time per qry (m:ss)	Prec 20	Time per qry (m:ss)	Prec 20
2 GB	INQ402	INQ414	0:41	0.387	0:23	0.389	0:20	0.324
20 GB	INQ402	INQ412	6:50	0.505	3:48	0.497	2:59	0.332

4.3 VLC analysis

The most striking result of the VLC experiments is that precision is far higher on the 20 GB corpus than on the 2 GB baseline corpus. This result is not unique to Inquery; every group participating in the VLC track had similar results. Its cause is unknown, although it may simply be that the larger database had more relevant documents.

A second result was that query time scaled linearly with the size of the database. This result was expected, because we used a version of Inquery that does not do any form of optimization.

An unofficial experiment tested the effects of *top-docs* optimization, in which each query term contributes only its best 1,000 documents to the ranked list. The top-docs optimization had minimal impact on precision while doubling the speed of document retrieval, which is consistent with published results [5].

Another unofficial experiment tested the effects of *top-docs-only* optimization, in which each query term contributes a score for only its best 1,000 documents. The top-docs-only optimization improved speed by another 13-21% (as compared with the top-docs optimization), but reduced precision by 17-33%. These results were a surprise; we expected more of an improvement in speed, and less of a loss in precision.

The timing experiments demonstrate that the current optimization techniques do not provide the speed necessary to run highly complex queries on a 20 GB database. The queries created for TREC Ad-hoc experiments contain an average of 99 terms and 31 query operators (primarily proximity and phrase operators) per query. Although effective, few people would wait 3-4 minutes for query results – even for very good results. A combination of more concise queries and improved optimization techniques are required for very large corpora.

5 Filtering track

Our goals for the Filtering track were to use InRoute, our document filtering system [8], for all of the experiments, and to use an incremental Rocchio algorithm [1] for the Adaptive Filtering experiments. These were modest goals, given our previous work. The only new work required was an algorithm to learn dissemination thresholds incrementally.

5.1 Filtering approach

The "batch-learned" experiments were of minimal interest to our group, because of their similarity to the Routing track. For example, the batch-learned profiles in all of our Filtering experiments were created with the same techniques used in the Routing track (described above). The filtering experiments merely used a more restricted set of corpus statistics and relevance judgements. The batch-learned dissemination thresholds were the "optimal" thresholds for the training data [2].

Our interest in the "batch-learned" experiments was confined to seeing the effects of different corpus statistics, and the effects of different evaluation metrics. Consequently, seven of our ten runs are quite similar.

The Adaptive Filtering experiments were the most interesting to us because of their similarity to "real world" environments. Each topic was converted automatically into an AdHoc query, using a subset of the techniques used in the AdHoc track (described above). The initial dissemination threshold was set low enough that matching on any query term would exceed the threshold.

During the training phase, if a document was selected for dissemination, InRoute was given that document's relevance judgement; unjudged documents were treated as not relevant. Profiles were modified using

Run ID	Profile Method	Threshold Method	Corpus Stats	Metric	Prec100	AvPrec
INQ415	Batch	Batch	FBIS 3,4	F1	0.1111	0.0499
INQ416	Batch	Batch	FBIS 3,4	F2	0.1705	0.0734
INQ417	Batch	Batch	TREC 1,2,3 +	F1	0.0746	0.0391
INQ418	Batch	Batch	TREC 1,2,3 +	F2	0.1417	0.0656
INQ419	Batch	Batch	FBIS 3,4	ASP	0.0087	0.0039
INQ420	Batch	Batch	TREC 1,2,3 +	ASP	0.0115	0.0046
INQ421	Online	Online	FBIS 3,4	N/A	0.2670	0.1683
INQ421c					0.3297	0.2074
INQ422	Online	Online	TREC 1,2,3 +	N/A	0.2924	0.1698
INQ422c					0.2817	0.1794
INQ423	Online	N/A	FBIS 3,4	Ranked	0.2668	0.2067
INQ423c					0.3270	0.2774
INQ424	Batch	N/A	FBIS 3,4	Ranked	0.2306	0.1525
INQ424c					0.2864	0.2075

Figure 4: Summary of the ten UMass Filtering runs. Run names postfixed with "c" are corrected versions of the official TREC submissions.

an incremental Rocchio algorithm [1]. Thresholds were modified to be halfway between the average relevant document score and the average nonrelevant document score.

Profiles and thresholds were "frozen" during the testing phase.

The three adaptive runs differ in the corpus statistics used, and the way in which they are evaluated.

Although 10 runs were submitted (Figure 4), the number of ideas tested was small.

- INQ415, INQ416, and INQ419 are identical *except* threshold learning; thresholds in these runs were "optimized" for different evaluation metrics (F1, F2, and ASP, respectively).
- INQ417, INQ418, and INQ420 are the same as INQ415, INQ416, and INQ419 *except* that a broader set of corpus statistics was used during filtering (TREC 1,2,3,+ instead of FBIS 3,4).
- INQ422 is the same as INQ421 *except* that a broader set of corpus statistics was used during filtering (TREC 1,2,3,+ instead of FBIS 3,4).
- INQ423 and INQ424 are the same as INQ421 and INQ415, but are evaluated as ranked runs.

The same "batch learned" profiles were used for runs INQ415 – INQ420, and INQ424; only the thresholds and corpus statistics differed among these runs. The "batch learned" profiles were learned using only FBIS 3 and FBIS 4 training data and corpus statistics.

5.2 Filtering results

The results are summarized in Figure 4. Several small errors were found and fixed after the official submissions, which led to improved results. The corrected runs are also shown in the table, with the suffix "c" added to the original run id.

5.3 Filtering analysis

Most of the batch-learned-profile experiments (INQ415-INQ420) produced poor results, due to poor selection of batch-learned thresholds. For example, the median number of documents disseminated by experiment INQ415 was 4. We have not yet done failure analysis to determine what caused the batch-learned thresholds to be so poor.

The one experiment that evaluted batch-learned-profiles using ranked retrieval (INQ424), instead of a dissemination threshold approach, produced results that were similar to Routing track experiments. This result was expected, because the experiment was essentially a Routing track experiment; the only differences for the Filtering track were a narrower set of corpus statistics (as required), and less accurate *idf* values for proximity operators (corrected in run INQ424c).

The experiments that tested adaptive learning methods were far more encouraging. Profiles learned adaptively (INQ421c-INQ423c) had precision and recall comparable to profiles learned with a batch method (INQ424c). Recall was lower when adaptively learned thresholds were applied (compare INQ421c to INQ423c), however the difference was smaller than expected (almost *any* threshold lowers recall). In these experiments, the adaptive methods of learning profiles and dissemination thresholds were quite effective.

Analysis of the experimental results is complicated by the fact that INQ421c and INQ422c did not find documents for every query, due to their use of thresholds. INQ421c found documents for only 36 queries, while INQ422c found documents for 46 queries. This result suggests that a broad set of corpus statistics is more effective than a narrow set, but one cannot draw strong conclusions from this one comparison.

It is difficult to draw conclusions about the threshold algorithm from these results. Clearly it did not cause profile-learning, Precision at 100 documents, or Average Precision to deteriorate substantially. However, a threshold of 0 produces similar results when there are many relevant documents. Further work is required to determine the effectiveness of the threshold algorithm.

Although many questions remain, the results from these experiments are encouraging. The adaptive results are comparable to batch results, even though proximity operators are not yet included in adaptive queries. (Proximity operators normally improve effectiveness significantly.) This result, by itself, is cause for optimism.

6 Chinese track

For TREC-6, we did not attempt any new processing of the queries or database for the Chinese track.

6.1 Chinese approach

The Chinese retrieval experiments are similar to the work done for TREC-5.

1. To allow for flexibility in segmentation at query time, each Chinese character is indexed as a term. Exceptions are made for characters making up numbers and the elements of dates which are indexed as a group.

2. Queries are made up of the title and description fields of the topics. They are automatically preprocessed to remove punctuation. These basic queries are then automatically segmented using the USEG segmenter, based upon hidden Markov models. Each segmented Chinese word is represented by a proximity operator which requires that the glyphs be immediately adjacent and in order. To compensate for possible segmenter errors, sequences of single characters are wrapped in a `#phrase` operator with the restriction that all glyphs be within a window of 25 terms. Each word in the description is weighted as a single term (weight 1.0) while isolated single terms are downweighted (weight 0.3). The whole title is weighted as a single term (weight 1.0).

3. The queries are expanded using Local Context Analysis (LCA). The basic query is used to retrieve the top-ranked passages for each topic. LCA is applied to extract expansion words from the top-ranked passages. An expansion word is a segmented word as defined by USEG. The segmenter is augmented with a name recognizer to reduce errors of name segmentation. The top 70 words from the top-ranked passages are added to the query. Each concept is assigned a weight in decreasing order. Word_i is assigned the weight $w_i = 1.0 - 0.9(i-1)/70$. Two runs are done. The first, INQ4ch1, extracts the expansion words from the 10 top-ranked passages retrieved and the second, INQ4ch2, from the 20 top-ranked passages retrieved. The expansion section of the final query is given twice the weight of the original query.

6.2 Chinese results

The following table summarizes our Chinese runs.

	Avg Prec	Prec @ 20	R-prec
title+desc, noseg	0.4785	0.7288	0.4831
title+desc, seg	0.4554	0.6827	0.4665
desc+seg	0.4209	0.6538	0.4324
title+seg	0.3743	0.5788	0.4088
INQ4ch1	0.5336	0.7654	0.5218
INQ4ch2	0.5223	0.7538	0.5137

6.3 Chinese analysis

It is surprising that the segmentation actually hurts the queries. We have not yet examined why this is true.

7 Cross-language IR (CLIR) track

The cross-language retrieval experiments focused on disambiguating translations of Spanish (source) queries to English (target). A parallel corpus of UN documents from 1988-1990, obtained from the LDC, was used in addition to POS tagging to disambiguate term translations. Phrases were translated via information extracted from the Collins Spanish-English machine readable dictionary (MRD). Local Context Analysis (LCA) was employed prior to and after query translation to reduce the effect of poor translations.

A more detailed discussion of some of the techniques used in this track was published recently.[4] Appendix A includes the CLIR Track Questionnaire.

7.1 CLIR approach

Query processing for the cross-language experiments begins with part-of-speech (POS) tagging using the MITRE POS tagger. As is the case with English queries, stop phrases are removed. With the exception of adjacent proper nouns which are treated as phrases, query and expansion terms terms in the source language are translated to the target language using the Collins MRD. The term translations are then disambiguated with the UN corpus. A more detailed description of query translation follows.

Each tagged query term is replaced with the source language equivalent term or terms that correspond to its part-of-speech. If there is no translation corresponding to a particular query term's tag, the translations for all parts-of-speech listed in the dictionary for that term are returned. There may be one or more ways to translate a given term. When more than one equivalent is returned, the best single term is chosen from this list via parallel corpus disambiguation.

Disambiguation proceeds in the following way. The top 100 Spanish documents are retrieved from the parallel UN corpus using the original Spanish query. The top 5000 terms based on Roccio ranking are extracted from the English UN documents that correspond to the top 100 Spanish documents. The translations of a query term are ranked by their weight in the list of 5000. The highest ranking equivalent is chosen as the "best" translation for that term. If more than one translation equivalent have the same rank, they are all chosen. If none of the equivalents are on the list, no disambiguation is performed and all equivalents are chosen.

Phrasal translations were performed using information on phrases and word usage contained in the Collins MRD. This allowed the replacement of a source phrase with its multi-term representation in the target language. When a phrase could not be defined using this information, it was translated word-by-word as described above.

Translated queries are then expanded using Local Context Analysis. When expanding, the top 50 concepts were added from the top 30 passages with multi-term concepts wrapped in the INQUERY #phrase operator with the restriction that all terms be found within a window of 25 terms. For example, #passage25(#phrase(president kurt waldheim)). Concepts were weighted with an infinder-like weighting scheme. The top concept was given a weight of 1.0 with all subsequent concepts down-weighted by $\frac{T-i-1}{T}$, where T is the total number of concepts and i is the rank of the current concept.

Two sets of queries were generated, one using only topic descriptions (INQxl2) and the other using both descriptions and titles (INQxl1). The original query translation and additional concepts were combined as described in the discussion of LCA (Section 1.3) with w_{lca} set to 1.0.

7.2 CLIR results

Two sets of results, INQxl1 and INQxl2, were submitted in the Cross-language track. Both sets were based on automatic processing of TREC topics CL1-CL25 into queries and automatic query expansion. The official results for 21 queries are summarized below. Table 3 compares effectiveness of English queries consisting of title plus description with queries INQxl1. Table 4 compares effectiveness of English description only queries with queries INQxl2. In both cases, the baseline English queries were expanded with the top 50 concepts from the top 30 documents.

Query Type	Precision			
	5 docs	30 docs	100 docs	Avg Prec (NI)
Desc	0.5429	0.4683	0.2814	0.3721
INQxl2	0.2095	0.1825	0.1167	0.1810 (-51.4)
INQxl2-fix	0.4000	0.3095	0.2043	0.2528 (-32.1)

Table 3: Results for title and description queries.

Query Type	Precision			
	At 5 docs	At 30 docs	At 100 docs	Ave Prec (NI)
Desc+Title	0.6000	0.4905	0.3081	0.4113
INQxl1	0.3048	0.2778	0.2010	0.2610 (-36.5)
INQxl1-fix	0.3619	0.3095	0.2019	0.2593 (-36.9)

Table 4: Results for description only queries.

Early analysis revealed programming errors which led to key query term translations being eliminated. For example, the pre-translation expansion term translations were not included in any query. We re-ran these experiments after eliminating the errors and the are shown in the third row of tables 3 and 4.

7.3 CLIR analysis

In the absence of complete relevance judgments, we are unable to perform an accurate analysis. However, we can say how these results compare to earlier work in cross-language retrieval. Cross-language retrieval via simple dictionary query translations [4, 3, 10, 12] tends to yield effectiveness which is 40-50% of monolingual retrieval effectiveness. Our cross-language description only query (INQxl2) results are consistent with this. Dictionary translations can be disambiguated via pre-translation and post-translation query expansion [4] or via part-of-speech and parallel corpus disambiguation [10], yielding cross-language effectiveness that is 70% of monolingual.

The TREC results are consistent with earlier results. However, we were surprised to find that pre-translation expansion alone was not particularly effective. We speculated that the overall effectiveness of the combined expansion method would improve if the effectiveness of the pre-translation expansion phase were improved. This turns out to be the case.

Table 5 shows representations of query 19 with both description and title. First is the original English, second the Spanish version, third the top 5 pre-translation expansion terms for the Spanish query, fourth the UN disambiguated translations of the expansion terms, and fifth the correct translations of the expansion terms. The disambiguation chooses the wrong translation about 20% of the time, shifting the query away from the correct context. Post-translation expansion may then pull in more unrelated concepts. If disambiguation is not used for expansion term translation, effectiveness of the pre-translation expansion increases as does the effectiveness of combining pre- and post-translation expansion. Table 6 shows an increase in effectiveness to 73% of monolingual when parallel corpus disambiguation is not used on the expansion term translations. Row one shows the original INQxl1 results and row two gives results for these queries without expansion-term corpus disambiguation. It is clear that although corpus disambiguation is effective, poorly disambiguated translations can have a large negative effect on performance.

The effect of each stage of the translation process as a percentage of monolingual average precision can be seen in table 7.

English	Wine. Is wine consumption production rising or decreasing world-wide?
Spanish	Vino. Está la producción consumo de vino creciendo o decreciendo a nivel mundial?
Exp. Terms	vino vinos consumo producción hule (bad term)
Dis. trans	party party consumption production rubber
Correct trans	wine wine consumption consumption n/a

Table 5: Query CL19

Query Type	Precision			
	At 5 docs	At 30 docs	At 100 docs	Ave Prec (NI)
INQxl1	0.3619	0.3095	0.2019	0.2593
INQxl1-no_dis	0.4095	0.3730	0.2424	0.3012 (+16.1)

Table 6: Precision at low recall and average precision for INQxl1 with and without corpus disambiguation of pre-translation expansion terms.

Query	Avg. Prec	%Monolingual
WBW	0.1570	38
WBW+Phr	0.1629	40
WBW+Dis	0.2099	51
WBW+Dis+Phr	0.2551	62
WBW+Dis+Phr+Pre	0.2454	60
WBW+Dis+Phr+Post	0.2864	70
WBW+Dis+Phr+Combined	0.2864	73

Table 7: Effect of translation steps as a percentage of monolingual average precision. WBW: word by word translation; Phr: phrase (proper nouns) recognition and translation; Dis: POS and UN corpus disambiguation; Pre: pre-translation expansion; Post: post-translation expansion; Combined: pre- and post- translation expansion.

8 Spoken Document Retrieval (SDR) track

Our efforts in this track compared runs on three databases: the human transcribed text, the provided recognized text, and text recognized by Dragon Systems on our behalf. In all cases, we used minimal query processing methods and two rounds of LCA to generate the queries.

8.1 SDR approach

Our SDR work utilized four sets of documents:

1. The LTT corpus provided by NIST. These are human-transcribed texts of the audio corpus. They provide the expected upper bound of performance.
2. The IBMSRT corpus, also provided by NIST. This corpus is the result of IBM's providing speech-recognized text for use by the entire SDR group. It is degraded text.
3. The DRAGON corpus, built by Dragon Systems, our partners in this track. This corpus is also degraded text. The method used by Dragon to create the text is provided below.
4. The Topic Detection and Tracking (TDT) corpus available via the Linguistic Data Consortium. This is a set of about 16,000 news stories from Reuters and CNN, covering July, 1994, through June, 1995. It was used in this track as a reliable (non-degraded text) corpus covering a similar time period as the test corpus.

The first three were test corpora and final queries were run against them for submission to NIST. The last corpus was used only during query construction.

For each test corpus, we created a 3-part query. The parts were:

1. The original query with stop-phrases removed and phrases identified as in the ad-hoc track (Section 2).
2. An LCA expansion of the original query using the TDT corpus. Up to 29 features were added from the TDT corpus. These were intended to provide additional features from a related corpus of high quality. (LCA expansion is described in Section 1.3.)
3. An LCA expansion of the original query using the test corpus (either LTT, IBMSRT, or DRAGON). Up to 29 features were added here, too. These were intended to expand the query based on the database to provide topical vocabulary.

The three parts were combined as a weighted sum:

```
#wsum(1.0 10.0 original-query
           2.0 test-LCA
          10.0 TDT-LCA )
```

Note that the expansion features from the test corpus were down-weighted relative to the other features. This was done because we felt that features extracted from a degraded database would be less reliable.

8.2 SDR speech recognition

The speech recognition component of our TREC SDR work (labeled "DRAGON" above) was accomplished by Dragon Systems. This section describes the process they used to transform the audio into text.

8.2.1 Acoustic models

The frontend that we are using has 36 features, namely 12 modified plp cepstra (including C0), and the corresponding first and second differences. Channel normalization is done within a given speaker's data.

The phone set that we are using is larger than we have used in the past: 51 phonemes (including silence) instead of the 43 phoneme set that we have used before. It is larger because certain vowels have stressed and unstressed versions, and it includes syllabic consonants.

We trained acoustic models using the first half of the HUB4 acoustic training corpus. We only used the first half so that we could use these models in the TREC SDR evaluation. This half of the data consists of about 34 hours of usable training material—however to start with, we trained only from speakers that had a minute or more of data in the first half. Overall, 27 hours of data distributed among 417 speakers satisfied this condition.

We used gender-independent models trained from a 24 hour subset of the WSJ si284 corpus to obtain initial alignments of the HUB4 data.

8.2.2 Clustering

In the TREC SDR evaluation we did not use the speaker side information, so we needed to develop a clustering algorithm that would group the data into clusters that corresponded to the actual speaker clusters.

To do the clustering, we use a k-means algorithm that uses the following distance measure of a segment s to a cluster c:

$$\mathrm{KL}(s, c+s) + \mathrm{KL}(c, c+s) + \mathrm{TimePen}(c, s)$$

where $\mathrm{KL}(a,b)$, the Kullbeck-Leibler distance, is the expectation under a (as the true hypothesis) of the logarithm of the ratio of the probability of the a distribution to the probability of the b distribution, and $\mathrm{TimePen}(a,b)$ is a linear function of the smallest time difference between a frame in a and a frame in b, truncated at a maximum value.

8.2.3 Language model

We used an interpolated language model consisting of two components:

1. A bigram language model trained from the first half of the acoustic training transcripts (roughly 400,000 words, with all bigrams kept).

2. A trigram language model trained from 62 million words of Journal Graphics transcriptions of broadcast news sources from the period January 1995 through April 1996 (kept all bigrams, but only trigrams that occurred three or more times). The Journal Graphics transcripts were processed to covert them from "written" text to "spoken" text.

Interpolation weights were trained from the 1996 HUB4 evaluation transcripts. The 56,000-word lexicon was constructed from three sources:

1. the 18,000 distinct words found in the first half of the HUB4 training data

2. the 19,000 most common new words found in the Journal Graphics data

3. the 19,000 most common new words found in 50 million words of newspaper data taken from the 1995 Philadelphia Inquirer.

8.3 SDR results and analysis

To illustrate the query processing methods, we consider Topic 3 in the SDR track. Words in quotation marks are phrases.

- *Original*: What is the difference between the old style classic cinemas and the new styles of cinema we have today?

- *Basic query processing*: difference "old style" old style classic cinemas new styles cinema

- *TDT expansion features*: frankenstein "film industry" "kenneth branagh" cinema film fad style lowrie "paris cinema" "fred fuchs" "francis ford coppola" "cinemas benefit" "century rendition" "century horror classic" "adrian wootton" "art form" prod. "mary shelley" casting technician "thai house" "peter humi" profit "robert deniro" popularity "margaret lowrie" helena hollywood image

- *IBM-recognized text expansion features*: heart yeltsin loom dollar "style rally" "russians dozen" "men mahal room hut" "m. men" "louisville ala" lerner "house canvassing" "ham men" "election spending" "economists yeltsin" "campaign team" "attitude moon" percent "v. broadcast" "soprano maria callas" "new line cinema" monitoring "mel gibson" janine "daniel m. t." movie "lou duva" equivalent news singapore

- *Dragon-recognized text expansion features*: years trent houses emission style graduate pandering nights negotiations cinema barrels awards kidney lott enemies "years industry" sander "houses emission" "g. o. p. fire brand set" wilderness tumor melting "majority leader trent lott" literature "cover story" dennis "house republicans" toronto soprano sequence

It is clear from the expansion features that the recognized text caused expansion with very poor, generally unrelated features.

The following table lists the number of topics (out of 49) where the known relevant item was ranked first by our system, and where it was found somewhere in the top 10 (including the first rank). Note that for the two topics that had two relevant documents (43 and 48), we always found those at ranks 1 and 2.

	LTT		IBMSRT		Dragon	
	top	top10	top	top10	top	top10
Basic	38	46	33	42	36	43
+TDT	35	45	25	42	34	44
+LCA	40	46	32	42	38	43
all	39	45	32	42	38	45

In the table, the rows correspond to basic query processing, adding the TDT expansion concepts, instead adding the expansion concepts from the database in question, and adding both sets of concepts. The colums correspond to the three collections: human-transcribed, machine transcribed for NIST, and Dragon's machine transciption.

Given the apparent quality of expansion concepts from the TDT and test corpora list above, it is surprising that adding the TDT concepts consistently hurt performance and adding the others often helped. However, Topic 3 may not be the ideal sample. The following lists three spectacular failures of the system:

1. In Topic 3, the known item was retrieved at rank 27 on the human transcribed corpus, and rank 209 on the Dragon run.

2. In Topic 42 (fashion in beach coverups), the relevant document was found at rank 24 (Dragon corpus). The TDT expansion added words that were vaguely on-point, but the Dragon expansion included oil refineries, coastlines, and wildlife refuges because of the word "beach."

3. In Topic 47 (the Valujet crash), the relevant document was found at rank 36 (Dragon corpus). This is primarily because although the TDT expansion included information about the Everglades, it focused on the sugar industry.

The errors in our system appear to be primarily the result of mistakes in query expansion—i.e., expanding the wrong word or the right words but in the wrong way—rather than because of limitations in the recognition of speech.

9 Interactive track

We designed a novel interface specifically for doing aspect oriented retrieval. This system had the following features:

- In order to save a document, it was necessary to drag it to an area reserved for aspects.
- Significant terms were extracted from documents grouped into an aspect to help the user in labelling an aspect.
- Color coded visual cues were provided to show a user if a document had been viewed before or not.
- A 3-D map was given to the user where documents with high similarity were placed close together.

Because the interface to our system was quite different from the control system, ZPRISE, and because it included two distinct visualizations (discussed below), we decided that even if a significant difference was found between our system and ZPRISE we would not know which part of the interface caused that difference. We then made two versions of our system: our full system ("AspInquery Plus") contained all the features listed above, and a more basic version ("AspInquery") that only used one of the visualizations (the only change to the code was commenting out a call to the constructor for the other visualization). If a large difference in performance was observed between the two systems we would then know what feature had caused it.

The work described below is discussed in more detail elsewhere.[7] Appendix B includes the protocol for one participant on Topic 1.

9.1 Interactive approach

As required by NIST, we ran ZPRISE as a control system; the two experimental systems were basic and extended versions of one program. The extended version ("AspInquery Plus") simply added a 3-D window to the basic system ("AspInquery") . Both versions use the well-known Inquery search engine. The core of our user interface has much in common with the ZPRISE interface, differing in two significant ways: ZPRISE displays the query terms contained in a document after the headline but our system does not, and our system color codes whether a document has been viewed but ZPRISE does not. Specifically, we write the headline information of a document in blue if it has not been viewed before, and purple if it has been seen. (This scheme was modeled after the default color scheme Web browsers use to show if a hypertext link has been followed or not.)

Both ZPRISE and our system accept plain text input for queries. Our system also supports a phrase operator, invoked by placing terms together within double quotes (e.g., "balanced budget"). The phrase operator increases the ranking assigned to documents where all terms in the phrase are found in close proximity. (In reality, our system supports the full syntax of Inquery, dozens of operators in all, but this is the only one we told participants about.)

The basic retrieval interface was extended with two additional windows: an "aspect window" to help the user collect and annotate found aspects, and (for AspInquery Plus) a 3-D visualization of document relationships.

9.1.1 Aspect Window

With a basic IR system, an analyst may be able to find the documents containing various aspects, but he or she has to use another window or a piece of paper to keep track of what has been found already. We implemented an "aspect window" tool to help with this task. The idea is to provide an area where documents on a particular aspect can be stored. To help label the information, statistical analysis of word and phrase occurrences is used to decide what terms and phrases are most distinctive about a document or set of documents in an aspect. We provided an area for the user to manually assign additional keywords or labels if needed.

Each area of the aspect window has a colored border, a text field at the top for entering a descriptive label, and an automatically generated list of the five noun phrases that most distinguish the group of documents assigned to this aspect from the remainder of the collection. The description field is solely for the user's convenience and need not be filled. If the user wants a description they can type or paste into it, or drag automatically generated phrases into it. The top of Figure 5 shows an example of the aspect window.

9.1.2 Visualization: 3-D Window

Another important step in the aspect oriented retrieval task is deciding (repeatedly) which document to look at next. Aspects represent different forms of relevance, and we believe that they will group together within the set of retrieved documents. AspInquery Plus compares retrieved documents in an extremely high-dimensional space (approximately 400,000 for this collection) where each dimension corresponds to a feature in the collection and the distance was measured by the sine of the angle between the vectors. That space was collapsed to three dimensions for visualization using a spring embedding algorithm.

In the 3-D visualization of the retrieved set, documents that have not been assigned to any aspect have the same blue/purple (read/unread) color scheme that is used in the main window. Documents in the 3-D window are persistent between queries: when new documents are retrieved they are colored light blue (light purple when read) and are placed in the 3-D window by the forces exerted from already placed documents. The bottom of Figure 5 shows five newly retrieved documents in light gray. It is easy to see that three of these documents fall into a group of two previously seen documents (upper right of figure) and the other

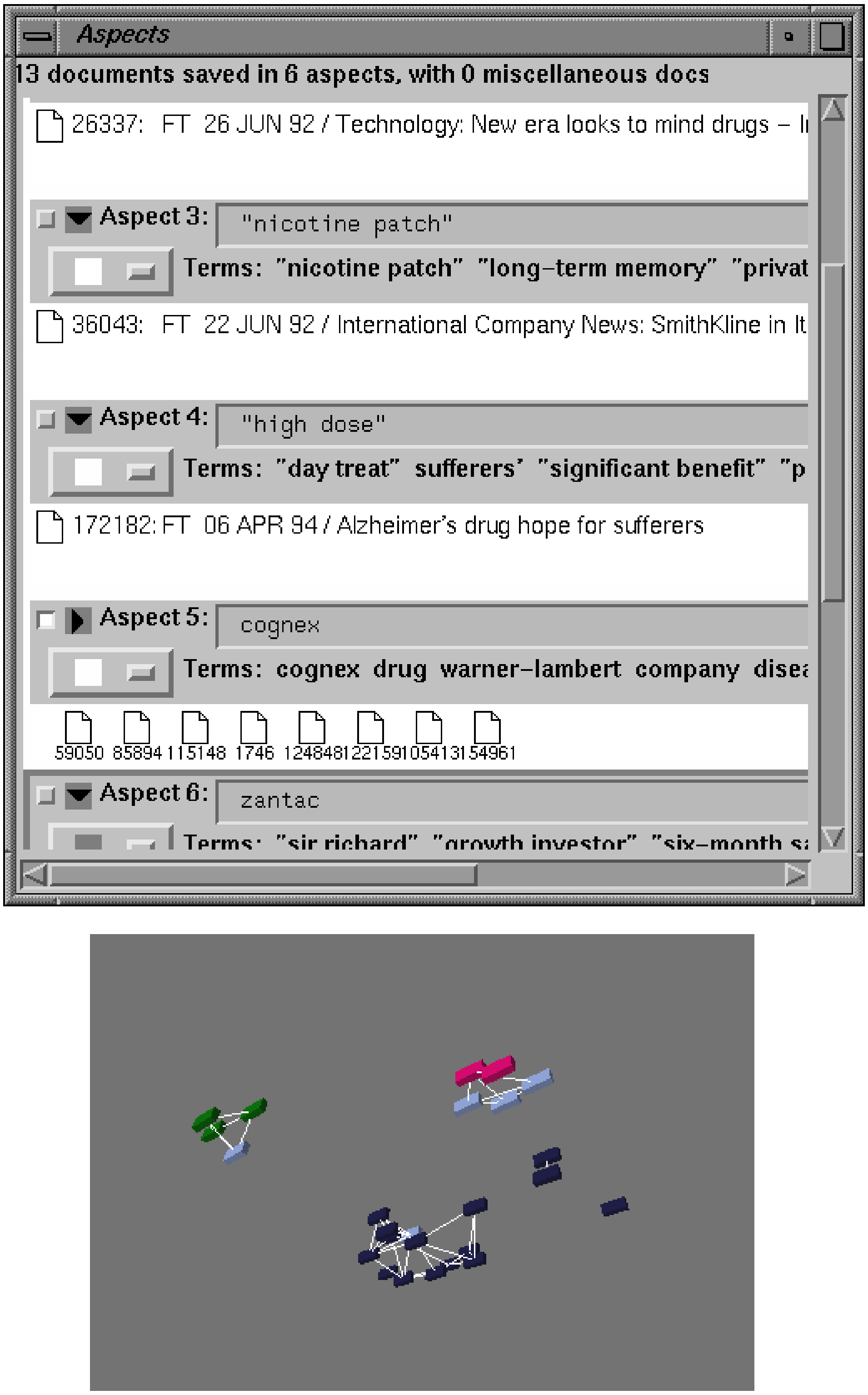

Figure 5: Visualizations provided by the interactive system. Aspect window for interactive system. The top box is the aspect window; the lower figure is the 3-D display

Group	Type	Control	Experimental	Size
1	General	ZP	AI	4
2	Librarian	ZP	AI	4
3	General	ZP	AI+	4
4	Librarian	ZP	AI+	4
5	General	AI	AI+	4

Table 8: Breakdown of participants by systems used for the interactive track

new documents fall into the small group in the upper left and the large group. An analyst who is under time pressure could use the 3-D display to decide that the unjudged document near that aspect is probably on the same aspect and so not worth examining. A retrieved document that is far from any already-marked aspect is more likely to be useful.

9.2 Participants for interactive task

We were interested in how librarians perform search tasks as compared to a more general user population. To that end, we recruited 20 participants: eight librarians and 12 general users. Table 8 shows the types of participants in and the systems used by the different groups in the experiment. Participants were told that the study would take about 3-1/2 hours and that they would be paid $35 if they completed it.

Seven of eight librarians were over 40; six of eight were women; all has very substantial experience with online searching, though had little experience with ranked lists or relevance feedback. The general participants were with one exception under 40; five of the twelve were women; they had moderate to no experience with on-line searching.

9.3 Interactive procedure

The experiment was run in the CIIR's usability laboratory. A "facilitator" was in the room with the participant all of the time except while the participant was doing the tutorials. The same person acted as facilitator for all participants except for the last two in group 5.

First, each participant filled out a questionnaire to give us basic demographic information (age, gender, degrees, general computer experience, experience with various types of searching, etc.). Each participant also took two standard psychometric tests from ETS: a test of verbal fluency (Controlled Associations, test FA-1), and a test of structural visualization (Paper Folding, test VZ-2).

Next, the participant was given a tutorial to learn one system, then they worked on the first three topics. After a short break they were given a tutorial on another system, then they worked on the other three topics. Each search had a 20-minute time limit, and the participant was instructed to stop the search if they had not finished in 20 minutes.

We gave each participant a piece of scratch paper before each search, and a short questionnaire after each. After all the searches were finished the participant was given a final questionnaire, and then "debriefed". The study was conducted single blind: the participants were not told until the debriefing which system was the control and which was the experimental system.

We ran each participant through the entire study in a single essentially continuous period of slightly over three to slightly over four hours, with no breaks longer than about 15 minutes.

9.4 Interactive results

The results are portrayed in Figure 6, a pair of graphs generated and provided by NIST.

9.5 Interactive analysis

Figure 7 shows the amount of variance that can be attributed to: topic, site, system, searcher, and random effects. This is based on a preliminary analysis of the data supplied by NIST of the 52 participants who

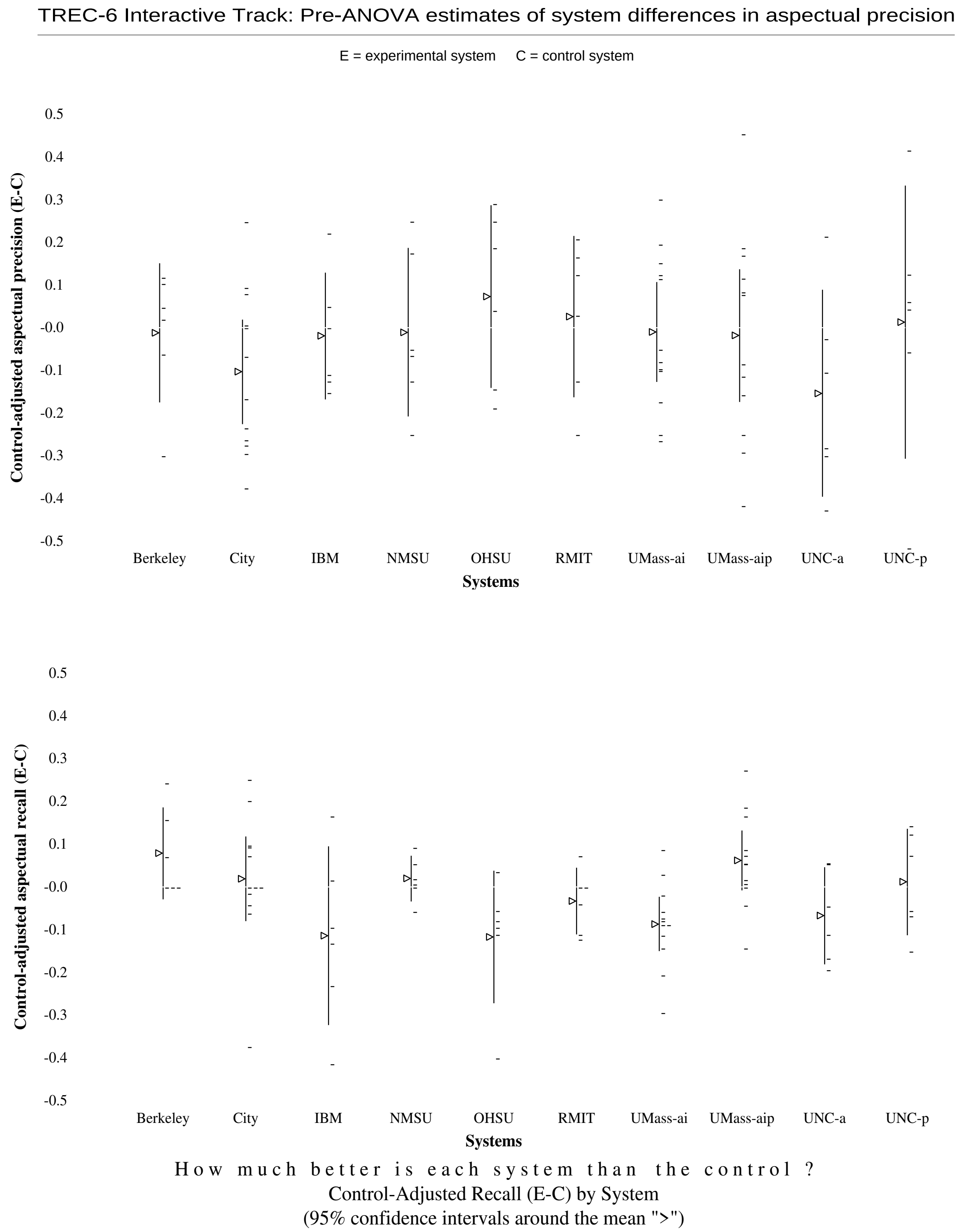

Figure 6: Graphic presentations of pre-ANOVA estimates of system differences via the control. Top graph describes precision; bottom graph, recall. (Provided by NIST.)

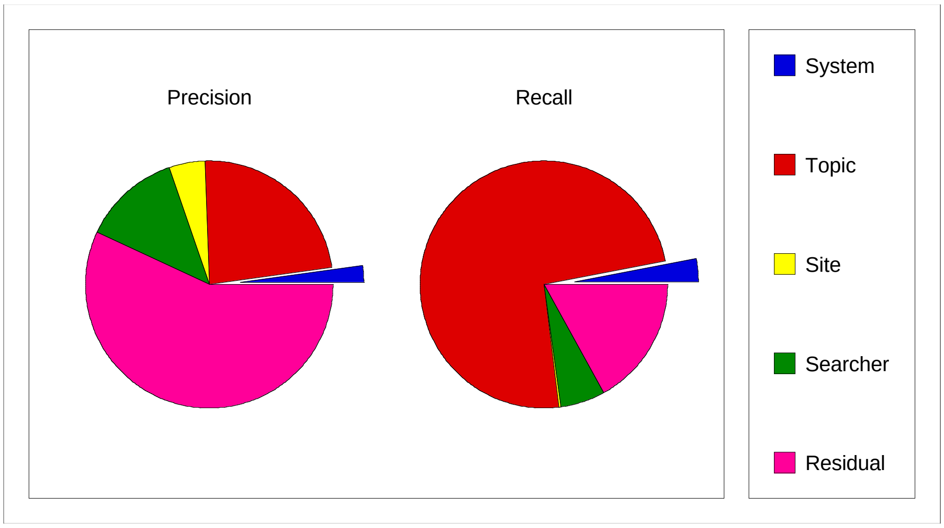

Figure 7: Sources of variance in interactive track, across all sites and all systems.

used ZPRISE as a control. The system differences are small relative to other sources of variation. Statistical analysis (ANOVA) has been performed by both NIST and CIIR, but whether or not statistically significant differences between systems was found depends on which test was used. In the following discussion "significance" claims are based on the tests showing significant differences between systems. Whether or not these differences really exist is discussed in the next subsection.

Figure 6 shows that most systems did not perform significantly different from the control. But at the CIIR *both* of our systems performed significantly different than the control, one worse and one better. (Part of the difference is we have a smaller confidence interval, as we ran 8 users per system and most sites ran 4 users per system.)

For the interactive task, the precision and recall scores are based on the relevance of documents that the searchers marked as being relevant. As a result, precision should be expected to be high. Precision would only be less than 1.0 if the searcher misunderstood a specific topic or made an error. The system effects should be small. Even if a system retrieved a very low precision set, the user must decide which documents are relevant. As can be seen from Figure 6, no system had a significant difference from ZPrise in precision.

For a set to have high aspectual recall, the system must retrieve documents representing all or most of the relevant aspects. The user must then judge those documents, and then save them. The recall score is then based on the recall of retrieved documents, the recall of the documents that are viewed, and the recall of the documents that are saved.

Our 2 systems differed from ZPRISE primarily in the interface presented to the user after a query was run. (The users were instructed in the use of the phrase operator, but most did not use it. Only 6 of the 16 participants in the main groups used it at all, and it was used on only 9 topics.) The aspect window had no features which would be expected to enhance recall. We expected no difference in recall between the AI system and ZPrise. As seen in the bottom graph in Figure 6, AI showed a significant drop in recall versus ZPrise. We are unsure of the reason for this drop. A possible explanation is that the interface is more complicated than the interface for ZPrise, and users had time trouble. We do not believe that this accounts for the difference.

Figure 8 shows the recall of the AI vs ZP for the 2 separate groups. The general group preferred AI over ZP 3 to 1, yet they did significantly worse with AI. The group of librarians preferred ZP over AI 4 to 0. They also did better with ZPrise than with AI, but by a smaller margin (AI outperforming ZPrise is within the confidence interval). Figure 9 shows the difference in time between the experimental system and the control system for the different groups. Group 1, which did significantly worse with AI, had no difference in time required between the two systems so time was clearly not a factor. Group 2 (librarians) did take

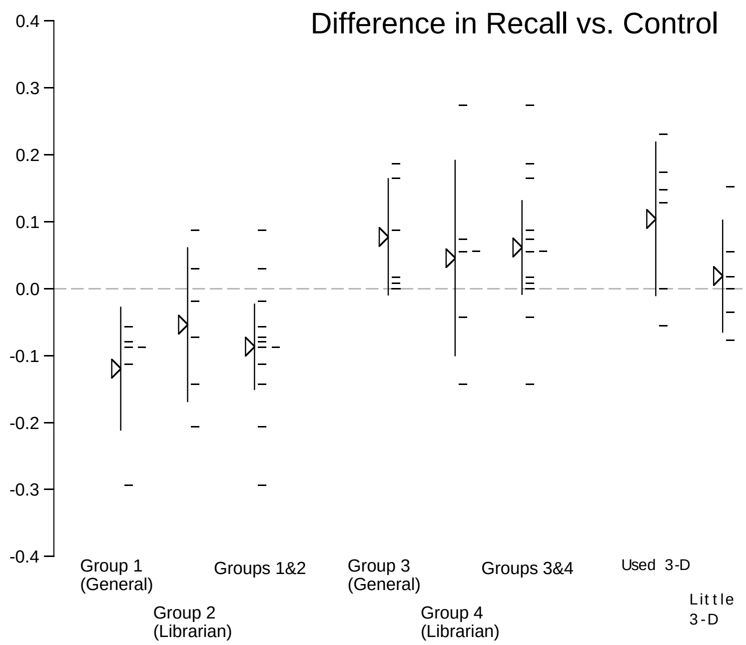

Figure 8: Recall broken down by system compared to control, and by groups of users within system.

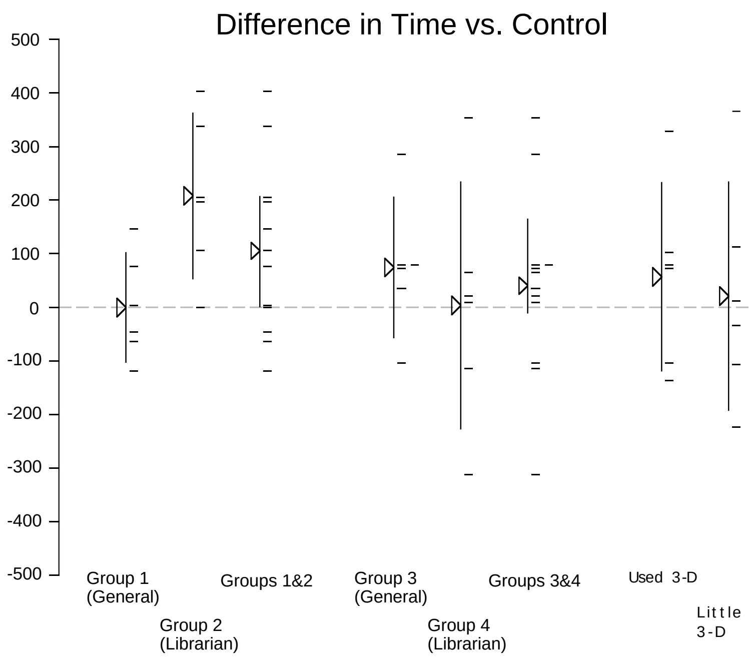

Figure 9: Difference in time spent on task broken down by system compared to control, and by groups of users within system.

Group	Number	Interactions
	11	4
General	12	181
	13	52
	15	0
	10	0
Librarian	14	10
	17	44
	19	143
	16	2
Final group	18	373
	20	148
	21	267

Table 9: Number of interactions with the 3-D window for 12 different users of the system.

signifantly longer with AI than with ZP (200 seconds on average). Time pressure may have been a factor with this group. However, this group did better on recall than group 1.

The other visualization used in our system, the 3-D window, was intended as a recall enhancing device. After a small number of documents have been viewed the 3-D map can be used to select documents that are likely to present new information, and can give better clues than a ranked list. This system showed a significant increase in recall versus ZPrise, and a very large increase in recall compared to AI. We expected an increase in recall with this interface, but we were surprised by the magnitude of the increase. We had learned from previous experience that users are often uncomfortable with 3-D interfaces and may not use them. We instrumented the 3-D window to record user interactions. Table 9 shows the number of interactions with the 3-D window for the different users. If the user ignored the window completely, the system he or she was using was the basic AI system with additional screen clutter. We would expect the results for users who did not use the 3-D window to be consistent with performance on the AI system. We divided the 8 participants into two groups, those who used the 3-D significantly (12, 13, 14, and 19), and those who didn't (11, 15, 10 and 17). (Participant 17 used the 3-D more than participant 14, but that breakdown did not complete the latin square design). The right-hard two bars of Figure 8 show the results for these groups. The group that used the 3-D had higher recall, and the group that didn't use 3-D had similar recall between AI+ and ZP.

9.6 Interactive methodology

NIST performed ANOVA results are reported elsewhere.NIST performed ANOVA on the averaged differences between the experimental systems and the control system (E-C) within each 2x2 Latin Square. The results show a significant difference between experimental systems across all sites, with p = .0133. Pairwise comparisons between systems were done using Tukey's Studentized Range. At p = 0.10, no significant differences were found pairwise between systems. The difference between the AI and AIP was 0.14825. For statistical significance at the 0.10 level, a difference between systems of 0.15033 was required. The obtained difference was 98.6

When the same analysis was performed on just the UMass data that compared a system against ZPRISE, the difference between the E-C data was .14825, for an F-value of 10.99, significant at p = 0.0035. When ANOVA was run on the model

y(i,j,k) = m + s(i) + t(j) + p(k) + e(i,j,k)

with y(i,j,k) being the recall value for searcher k using system i on topic j, we obtained an F value of 3.90, significant at p = 0.0245. The contrast between experimental systems AI and AIP showed a sum of squares of 0.13187 out of a total system based sum of squares of 0.13564.

The ANOVA performed by NIST showed the two UMass systems barely missing significance. The same analysis performed on just UMass data, and a different ANOVA on UMass data both show significance.

One of the hopes of the interactive track is that comparing systems against a common control will provide

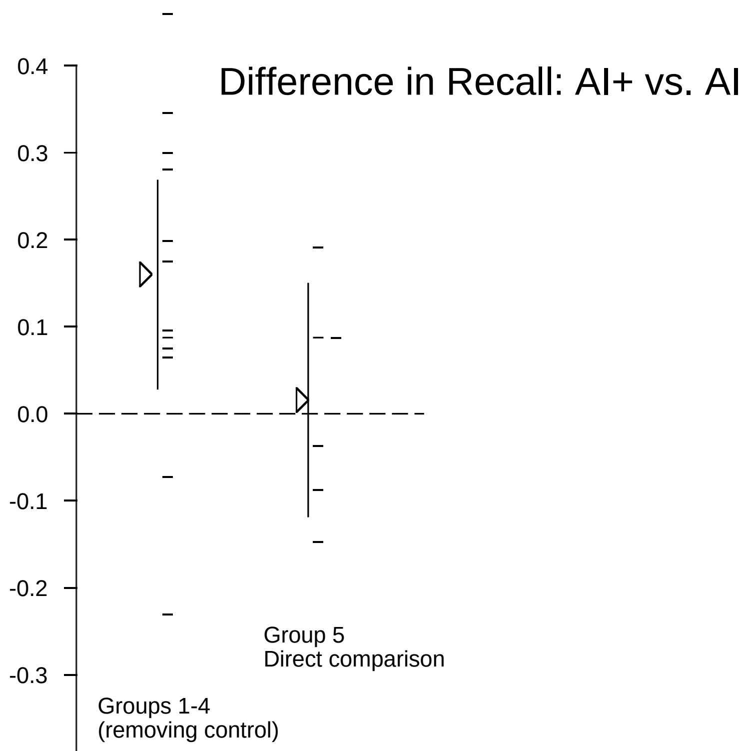

Figure 10: Difference in recall using AspInquery Plus as compared to AspInquery.

the same information as comparing two systems directly against each other. The pre-experiment was designed to validate this approach, but the results are inconclusive. We ran our two experimental systems directly against each other. From the results seen with ZPRISE as a control, we would predict that a significant difference in recall would be observed between the two systems. We did not obtain this result. We found that AI+ outperformed AI in recall with an average value of 0.0156, instead of the 0.14825 value given by the earlier experiment. These results are shown in Figure 10. ANOVA on the direct comparison showed an F value of 0.09, which is not significant. These results, combined with the inconclusive results in the preexperiment, raise questions about the validity of the approach taken in the interactive track.

ANOVA of the results on all five groups of participants showed a difference between systems with an F value of 2.49, $p = 0.089$.

9.7 Interactive conclusions

The system effects were observed with both librarians and a general population. The effects were attenuated on librarians.

Our two systems were much more like each other than they were like the control, but we obtained opposite effects. Since the only difference between the two systems was the 3-D window, we can conclude that providing a graphical display of document similarities as an alternative interface to a ranked list enhances recall in an interactive setting.

Analysis of Variance was performed on our data in several ways, and we obtained varying results. It appears that there is a (marginally) significant difference between our systems, but it is only apparent when measured against a control and is not apparent in direct comparisons. This raises questions about the assumptions and methodology used in the interactive track.

Acknowledgements

We thank Margie Connell, Aiqun Du, Victor Lavrenko, Anton Leouski, Daniella Malin, Darren Mas, Michael Scudder, and Kamal Souccar of the CIIR, as well as Steven Wegmann of Dragon Systems for their assistance in the work described here.

This study is based on research support by several grants: the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623; the National Science Foundation under grant number IRI-9619117; and the NSF Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst.

Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

[1] J. Allan. Incremental relevance feedback. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 270–278, Zurich, 1996. Association for Computing Machinery.

[2] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu. INQUERY at TREC-5. In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*. National Institute of Standards and Technology Special Publication, (in press).

[3] Lisa Ballesteros and W. Bruce Croft. Dictionary-based methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pages 791–801, 1996.

[4] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, Philadelphia, 1997. Association for Computing Machinery.

[5] E. W. Brown. Fast evaluation of structured queries for information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, pages 30–38, Seattle, 1995. Association for Computing Machinery.

[6] Chris Buckley and Gerard Salton. Optimization of relevance feedback weights. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, pages 351–357, Seattle, Washington, July 1995. ACM.

[7] Don Byrd, Russell Swan, and James Allan. TREC-6 interactive track report, part 1: Experimental procedure and initial results. Technical Report IR-117, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, November 1997.

[8] J. P. Callan. Document filtering with inference networks. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 262–269, Zurich, 1996. Association for Computing Machinery.

[9] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[10] Mark W. Davis and William C. Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval*, pages 92–98, 1997.

[11] Warren R. Greiff, W. Bruce Croft, and Howard Turtle. Computationally tractable probabilistic modeling of boolean operators. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 119–128, Philadelphia, 1997. Association for Computing Machinery.

[12] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 49–57, 1996.

[13] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 4–11, Zurich, 1996. Association for Computing Machinery.

A CLIR Track Questionnaire

A.1 OVERALL APPROACH:

A.1.1 What basic approach do you take to cross-language retrieval?

[x] Query Translation
[ ] Document Translation
[ ] Other, ______________________

A.1.2 Were manual translations of the original NIST topics used as a starting point for any of your cross-language runs?

[ ] No
[x] Yes, translated by a native spanish speaker then submitted to trec

A.1.3 Were the automatically translated (Logos MT) documents used for any of your cross-language runs?

[x] No
[ ] Yes, ______________________

A.1.4 Were the automatically translated (Logos MT) topics used for any of your cross-language runs?

[x] No
[ ] Yes, ______________________

A.2 MANUAL QUERY FORMULATION:

A.2.1 If query formulation involved manual effort, how fluent was the user in the source (query) language?

[ ] ______________________

A.2.2 If query formulation involved manual effort, how fluent was the user in the target (document) language?

[ ] ______________________

A.3 USE OF MANUALLY GENERATED DATA RESOURCES:

A.3.1 What kind of manually generated data resources were used?

[x] Dictionaries
[ ] Thesauri
[ ] Part-of-speech Lists
[x] Other, UN aligned corpus

A.3.2 Were they generated with information retrieval in mind or were they taken from related fields?

[ ] Information Retrieval
[ ] Machine Translation
[ ] Linguistic Research
[x] General Purpose Dictionaries
[ ] Other, ______________________

A.3.3 Were they specifically tuned for the data being searched (ie. with special terminology) or general-purpose?

[ ] Tuned for data; Please specify ____________________
[x] General purpose

A.3.4 What amount of work was involved in adapting them for use in your information retrieval system.

[ ] None
[x] involved cleaning mark-up meant for human users

A.3.5 Size

[ ] Collins about 50k______ entries
[ ] UN data: 500 MBytes

A.3.6 Availability? - Please also provide sources/references!

[ ] Commercial
[x] Proprietary, Collins spanish-english MRD
[ ] Free
[x] Other, UN data from the LDC

A.4 USE OF AUTOMATICALLY GENERATED DATA RESOURCES:

A.4.1 Form of the automatically constructed data resources?

[ ] Lexicon
[x] Thesaurus
[ ] Similarity matrix
[x] Other, phrase dictionary of word usage and phrasal information from dictionary

A.4.2 What sort of training data was used to construct them?

[x] Same data as used for searches, AP database
[x] Similar data as used for searches, El Norte collection
[ ] Other data, ____________________

A.4.3 Size

[x] Sp. Th: 112k, Phr. Dict:48k entries
[x] Eng. Th: 287, Sp. Th: 74, Phr. Dict: 6 MBytes

A.4.4 Was there any manual clean-up involved in the construction process?

[ ] Yes, ____________________
[x] No

A.4.5 Rough resource estimates for building the data resources (ie. an indicator of the computational complexity of the process).

[x] 200MB/hour (co-occurrence thesaurus)
[x] 120 MB memory used per 1 Gig. data
[x] about 2x collection size temporary disk space

A.5 GENERAL

A.5.1 How dependent is the system on the data resources used? Could they easily be replaced if better sources were available?

[ ] Very dependent, ______________________
[ ] Somewhat dependent, ______________________
[x] Easily replacable, ______________________
[ ] Don't know

A.5.2 Would the approach used potentially benefit if there were better data resources (e.g. bigger dictionary or more/better aligned texts for training) available for tests?

[x] Yes, a lot, e.g. specialized dictionaries
[ ] Yes, somewhat, ______________________
[ ] No, not significantly, ______________________
[ ] Don't know

A.5.3 Would the approach used potentially suffer a lot if similar data resources of lesser quality (noisier dictionary, wrong domain of terminology) were used as a replacement?

[ ] Yes a lot, ______________________
[x] Yes, somewhat, ______________________
[ ] No, not significantly, ______________________
[ ] Don't know

A.5.4 Are similar resources available for other languages than those used?

[x] Yes, ______________________
[ ] No

B TREC Interactive Track Protocol Log

The following is the log of the interaction for Participant 13, Topic 1 (326i). Spoken words are shown in italics. "U:" (for "User") precedes remarks and actions by the participant; "F:" precedes remarks by the Facilitator. Times are shown as "n s" for n seconds from the start of the session.

Time set at Tue Aug 12 09:24:57 1997
query is ferry sinking casualties
Query is: ferry sinking casualties
bl->term_freq = 0, default_belief = 0.400000, totalhits = 2932
bl->doc_cnt = 20
24 s Tue Aug 12 09:25:21 1997
Number of docs found is 20
1: 177935: FT943-312: FT 30 SEP 94 / Ferries in six 'near accidents': Finland and Sweden order checks after Estonia sinking
2: 205199: FT944-15661: FT 17 OCT 94 / World News in Brief: Bangladesh ferry sinks
3: 174281: FT943-178: FT 30 SEP 94 / Leading Article: Defying the cruel sea
4: 204595: FT944-15057: FT 20 OCT 94 / Improved ferry safety urged
5: 194241: FT944-5773: FT 02 DEC 94 / World News in Brief: Manila ferry sinks
6: 200503: FT944-11367: FT 07 NOV 94 / Pounds 45m car-ferry research planned
7: 199238: FT944-10102: FT 12 NOV 94 / Tighter ferry rules proposed
8: 208111: FT944-18217: FT 05 OCT 94 / World News in Brief: Check on ferries ordered
9: 200184: FT944-11048: FT 08 NOV 94 / Bow doors faulty on 33% of ferries using UK ports: Government to increase safety checks on vessels
10: 208769: FT944-18875: FT 01 OCT 94 / What future for the ferry?: Questions raised by the Baltic tragedy
11: 178166: FT943-543: FT 29 SEP 94 / Bow doors leak reported after 800 die in Baltic ferry sinking
12: 193552: FT944-5084: FT 06 DEC 94 / Ro-ro ferry study agreed
13: 75524: FT931-8485: FT 19 FEB 93 / Crowded ferry sinks off Haiti
14: 199245: FT944-10109: FT 12 NOV 94 / Tighter ferry rules proposed
15: 193716: FT944-5248: FT 05 DEC 94 / Sea safety review focuses on ferries
16: 178159: FT943-536: FT 29 SEP 94 / Safety rules that failed the Estonia: It was a modern ship, well maintained and partly Swedish owned. But are even the best ro-ro ferries vulnerable?
17: 177939: FT943-316: FT 30 SEP 94 / Ferries face calls for safety curbs: Estonia disaster brings reports of other 'near accidents'
18: 199989: FT944-10853: FT 09 NOV 94 / Eurotunnel hits at government on ferry safety
19: 39232: FT923-4546: FT 05 SEP 92 / Swan wins order for Tyne ferry
20: 39203: FT923-4517: FT 05 SEP 92 / Ferry order for Tyne yard
38 s, Reading doc 177935:FT943-312, click from main win, time Tue Aug 12 09:25:35 1997
0:42 U: *OK, so my first article is about a ferry that sank and 900 people died.*
71 s, Reading doc 177935:FT943-312, click from main win, time Tue Aug 12 09:26:08 1997
Doc number 177935 added to aspect 0
Aspect # 0, auto terms are estonia "estonia sink" forsberg "estonia disaster" "bow section"
No user supplied text
1:18 U: *This article describes six incidents.*
1:30 F: *Six incidents of ferry sinkings?*
U: *Right.*
1:48 U: *Oh, talks about six near accidents and it describes one that actually happened.*
148 s, Reading doc 205199:FT944-15661, click from main win, time Tue Aug 12 09:27:25 1997
Doc number 205199 added to aspect 1
2:30 U: *This is a brief article about 400 people that dies in Bangladesh so we'll save that, and the name ... we'll name that one Bangladesh.*
Aspect # 1, auto terms are "ferry sink" "ferry disaster" "wedding party" "high sea" bangladesh
U: drags "bangladesh" into label area
197 s, Reading doc 174281:FT943-178, click from main win, time Tue Aug 12 09:28:14 1997
3:18 U: *Here's another article about the Estonia incident. It's a repeat so I don't need to save that, or should I save that also under*
F: *No, there's no need, there's no point to saving additional ones.*

224 s, Reading doc 204595:FT944-15057, click from main win, time Tue Aug 12 09:28:41 1997
3:45 U: *This one just talks about the Estonia again, so we don't need that.*
232 s, Reading doc 194241:FT944-5773, click from main win, time Tue Aug 12 09:28:49 1997
3:53 U: *And the fifth one is in Manila, 480 people were on it, 275 were rescued, and they're still picking up survivors, so you can probably assume 100 people died, so go ahead and save it.*
Doc number 194241 added to aspect 2
Aspect # 2, auto terms are "ferry sink" "cargo ship" manila sink survivor
U: drags "manila" into label area
267 s, Reading doc 200503:FT944-11367, click from main win, time Tue Aug 12 09:29:24 1997
272 s, Reading doc 199238:FT944-10102, click from main win, time Tue Aug 12 09:29:29 1997
4:34 U: *A lot of these keep talking about tighter regulations due to the sinking of the Estonia.*
288 s, Reading doc 200184:FT944-11048, click from main win, time Tue Aug 12 09:29:45 1997
Doc number 200184 added to aspect 3
Aspect # 3, auto terms are "bow door" "marine safety agent" ferry "safety agent" "dr mawhinney"
4:57 U: *Here's one that briefly mentions a ship...*
5:09 U: *This one again is more about safety regulations, but it briefly mentions a ship that had 193 casualties, so I guess I'll type in my own word since the one I want isn't in there.*
U: types " Herald of Free Enterprise" in label area.
5:38 U: *I'm trying to name all these either by the name of the ship or where it happened.*
U: moves controls on 3-D window and alters view several times.
5:54 U: *I'm trying to see if I can use the 3-D to help me out*
F: *I'm sorry, you couldn't couldn't what... You're trying to see if*
U: *I'm trying to see if I can use this to give me ... I'm assuming that these are supposed to show articles in the connecting blocks that are more relevant*
F: *That are more similar to each other*
U: *More similar.*
407 s, Reading doc 208769:FT944-18875, click from main win, time Tue Aug 12 09:31:44 1997
421 s, Reading doc 178166:FT943-543, click from main win, time Tue Aug 12 09:31:59 1997
448 s, Reading doc 193552:FT944-5084, click from main win, time Tue Aug 12 09:32:25 1997
454 s, Reading doc 75524:FT931-8485, click from main win, time Tue Aug 12 09:32:31 1997
Doc number 75524 added to aspect 4
Aspect # 4, auto terms are "ferry sink" port-au-prince neptune haiti "product centre"
U: drags "neptune" into label area
485 s, Reading doc 199245:FT944-10109, click from main win, time Tue Aug 12 09:33:02 1997
491 s, Reading doc 193716:FT944-5248, click from main win, time Tue Aug 12 09:33:08 1997
493 s, Reading doc 177939:FT943-316, click from main win, time Tue Aug 12 09:33:10 1997
500 s, Reading doc 199989:FT944-10853, click from main win, time Tue Aug 12 09:33:17 1997
505 s, Reading doc 39232:FT923-4546, click from main win, time Tue Aug 12 09:33:22 1997
512 s, Reading doc 39203:FT923-4517, click from main win, time Tue Aug 12 09:33:29 1997
8:32 U: *So I went through all 20 of the articles. For the most part I'd say all but probably 3 or 4 talked about accidents with over 100 casualties, so should I try a new search?*
F: *It's up to you. You have plenty of time.*
U: *You mean try a different wording of it?*
F: *It's up to you.*
9:20 F: *You could also try raising the max docs.*
U: *OK.*
query is ferry sinking casualties
Query is: ferry sinking casualties
bl->term_freq = 0, default_belief = 0.400000, totalhits = 2932
bl->doc_cnt = 40
574 s Tue Aug 12 09:34:31 1997
Number of docs found is 40
1: 177935: FT943-312: FT 30 SEP 94 / Ferries in six 'near accidents': Finland and Sweden order checks after Estonia sinking

2: 205199: FT944-15661: FT 17 OCT 94 / World News in Brief: Bangladesh ferry sinks
3: 174281: FT943-178: FT 30 SEP 94 / Leading Article: Defying the cruel sea
4: 204595: FT944-15057: FT 20 OCT 94 / Improved ferry safety urged
5: 194241: FT944-5773: FT 02 DEC 94 / World News in Brief: Manila ferry sinks
6: 200503: FT944-11367: FT 07 NOV 94 / Pounds 45m car-ferry research planned
7: 199238: FT944-10102: FT 12 NOV 94 / Tighter ferry rules proposed
8: 208111: FT944-18217: FT 05 OCT 94 / World News in Brief: Check on ferries ordered
9: 200184: FT944-11048: FT 08 NOV 94 / Bow doors faulty on 33% of ferries using UK ports: Government to increase safety checks on vessels
10: 208769: FT944-18875: FT 01 OCT 94 / What future for the ferry?: Questions raised by the Baltic tragedy
11: 178166: FT943-543: FT 29 SEP 94 / Bow doors leak reported after 800 die in Baltic ferry sinking
12: 193552: FT944-5084: FT 06 DEC 94 / Ro-ro ferry study agreed
13: 75524: FT931-8485: FT 19 FEB 93 / Crowded ferry sinks off Haiti
14: 199245: FT944-10109: FT 12 NOV 94 / Tighter ferry rules proposed
15: 193716: FT944-5248: FT 05 DEC 94 / Sea safety review focuses on ferries
16: 178159: FT943-536: FT 29 SEP 94 / Safety rules that failed the Estonia: It was a modern ship, well maintained and partly Swedish owned. But are even the best ro-ro ferries vulnerable?
17: 177939: FT943-316: FT 30 SEP 94 / Ferries face calls for safety curbs: Estonia disaster brings reports of other 'near accidents'
18: 199989: FT944-10853: FT 09 NOV 94 / Eurotunnel hits at government on ferry safety
19: 39232: FT923-4546: FT 05 SEP 92 / Swan wins order for Tyne ferry
20: 39203: FT923-4517: FT 05 SEP 92 / Ferry order for Tyne yard
21: 207852: FT944-17958: FT 05 OCT 94 / Finns order ro-ro bow doors welded shut
22: 169325: FT942-14757: FT 19 APR 94 / Letters to the Editor: Channel control overdue
23: 207851: FT944-17957: FT 05 OCT 94 / UN maritime agency panel to review safety: A look at action prompted by the Baltic ferry disaster
24: 208393: FT944-18499: FT 04 OCT 94 / Baltic ferry operators to weld bow doors shut: Safety move follows confirmation of cause of Estonia disaster
25: 208098: FT944-18204: FT 05 OCT 94 / Maritime agency in safety plan
26: 208402: FT944-18508: FT 04 OCT 94 / Estonia's bow doors were torn off in heavy storm: Video of sunken ferry shows how water flooded car deck
27: 204681: FT944-15143: FT 19 OCT 94 / Estonia's missing bow door located
28: 200149: FT944-11013: FT 08 NOV 94 / International Company News: Heavy loss in US pushes Trygg-Hansa into the red - Swedish insurer posts SKr813m deficit at nine months
29: 141158: FT941-5434: FT 07 MAR 94 / Freight companies to shun Channel tunnel
30: 208832: FT944-18938: FT 01 OCT 94 / UN agency orders ferry probe: Estonia's bow doors may have been torn off in storm, Swedish authorities say
31: 178158: FT943-535: FT 29 SEP 94 / Tragedy leaves Swedes in shock
32: 171243: FT942-16675: FT 08 APR 94 / Survey of East Kent (7): Pain amid the gain - The ferries fight back
33: 195783: FT944-6974: FT 26 NOV 94 / Thinking the unsinkable: The modern parallels exposed by an exhibition about the Titanic, which sank in 1912
34: 205276: FT944-15738: FT 17 OCT 94 / Company News This Week: Departure delays leave investors counting the cost - Eurotunnel
35: 137406: FT934-1954: FT 16 DEC 93 / Technology: Ships bridge the danger gap - Andrew Fisher concludes a series on transport safety with an investigation into innovations that may help prevent sea disasters and give clues to their causes
36: 127095: FT934-8445: FT 16 NOV 93 / Corporate bankruptcies increase as demand sinks
37: 1655: FT911-4602: FT 18 APR 91 / MMC to investigate Isle of Wight ferries
38: 206611: FT944-1600: FT 19 DEC 94 / Survey of Sweden (14): A remarkable comeback - Profile: Stena Line
39: 26988: FT922-7334: FT 19 MAY 92 / World Trade News: Denmark-Sweden ferry link-up is agreed
40: 119826: FT933-1606: FT 23 SEP 93 / Ferry operator in link with Belgium
U: Several 3-D interactions
Reading doc 200503: FT944-11367, click from 3-D window
642 s, Reading doc 207852:FT944-17958, click from main win, time Tue Aug 12 09:35:39 1997
652 s, Reading doc 169325:FT942-14757, click from main win, time Tue Aug 12 09:35:49 1997
661 s, Reading doc 207851:FT944-17957, click from main win, time Tue Aug 12 09:35:58 1997
11:11 U: *So I increased the maxdocs from 20 to 40, and most of the later articles don't seem to really have much relevant information. Either they're talking*

about the Estonia or they're just talking about general safety regulations.
678 s, Reading doc 208393:FT944-18499, click from main win, time Tue Aug 12 09:36:15 1997
701 s, Reading doc 208098:FT944-18204, click from main win, time Tue Aug 12 09:36:38 1997
11:45 U: *I'm guessing that's why there's this big network here.* (Points to large cluster of documents in 3-D viewer.) *A lot of them are talking about the Estonia so I think they're all related in that sense.*
723 s, Reading doc 208402:FT944-18508, click from main win, time Tue Aug 12 09:37:00 1997
734 s, Reading doc 204681:FT944-15143, click from main win, time Tue Aug 12 09:37:11 1997
737 s, Reading doc 200149:FT944-11013, click from main win, time Tue Aug 12 09:37:14 1997
744 s, Reading doc 141158:FT941-5434, click from main win, time Tue Aug 12 09:37:21 1997
12:32 U: *Yeah this is really starting to get ... My query is "ferry sinking", and in this article the word "sink" only appears once, and it doesn't have anything to do with ferries, and there's nothing about casualties so it looks like we're getting farther and farther away from anything relevant. You can see that over here, we're moving further away from this point.* (Points to several documents in 3-D view)
775 s, Reading doc 208832:FT944-18938, click from main win, time Tue Aug 12 09:37:52 1997
800 s, Reading doc 171243:FT942-16675, click from main win, time Tue Aug 12 09:38:17 1997
813 s, Reading doc 195783:FT944-6974, click from main win, time Tue Aug 12 09:38:30 1997
822 s, Reading doc 205276:FT944-15738, click from main win, time Tue Aug 12 09:38:39 1997
827 s, Reading doc 137406:FT934-1954, click from main win, time Tue Aug 12 09:38:44 1997
13:49 U: *OK, I've found a new one.*
Doc number 137406 added to aspect 5
Aspect # 5, auto terms are moby imo vessel livorno ship
U: drags "livorno" into label area
14:06 U: *This is the first new article I've found in the last 20 I've looked at.*
868 s, Reading doc 127095:FT934-8445, click from main win, time Tue Aug 12 09:39:25 1997
872 s, Reading doc 1655:FT911-4602, click from main win, time Tue Aug 12 09:39:29 1997
876 s, Reading doc 206611:FT944-1600, click from main win, time Tue Aug 12 09:39:33 1997
884 s, Reading doc 26988:FT922-7334, click from main win, time Tue Aug 12 09:39:41 1997
892 s, Reading doc 119826:FT933-1606, click from main win, time Tue Aug 12 09:39:49 1997
15:00 F: *You have five minutes.*
15:23 U: *We'll try searching for ferry and accidents.*
query is ferry accident
Query is: ferry accident
bl->term_freq = 0, default_belief = 0.400000, totalhits = 1978
bl->doc_cnt = 40
932 s Tue Aug 12 09:40:29 1997
Number of docs found is 40
1: 174533: FT943-3295: FT 15 SEP 94 / Inquiry starts after six die in ferry walkway collapse
2: 42744: FT923-7671: FT 15 AUG 92 / Deaths ferry to be withdrawn
3: 149044: FT941-12581: FT 29 JAN 94 / Accident halts ferry services
4: 72637: FT931-5947: FT 03 MAR 93 / World News in Brief: Congo ferry toll rises to 146
5: 177935: FT943-312: FT 30 SEP 94 / Ferries in six 'near accidents': Finland and Sweden order checks after Estonia sinking
6: 186187: FT943-1246: FT 26 SEP 94 / World News in Brief: 16 injured in lifeboat accident
7: 208393: FT944-18499: FT 04 OCT 94 / Baltic ferry operators to weld bow doors shut: Safety move follows confirmation of cause of Estonia disaster
8: 208402: FT944-18508: FT 04 OCT 94 / Estonia's bow doors were torn off in heavy storm: Video of sunken ferry shows how water flooded car deck
9: 9804: FT921-686: FT 27 MAR 92 / Crash probe finds 'no abnormality'
10: 174478: FT943-3240: FT 15 SEP 94 / Investigators widen probe on ferry walkway collapse
11: 186180: FT943-1239: FT 26 SEP 94 / World News in Brief: 16 injured in lifeboat accident
12: 207852: FT944-17958: FT 05 OCT 94 / Finns order ro-ro bow doors welded shut
13: 177939: FT943-316: FT 30 SEP 94 / Ferries face calls for safety curbs: Estonia disaster brings reports of other 'near accidents'
14: 201958: FT944-12822: FT 31 OCT 94 / Business Travel: In S Korea, it is better to arrive ..
15: 208769: FT944-18875: FT 01 OCT 94 / What future for the ferry?: Questions raised by the Baltic tragedy
16: 14222: FT921-11074: FT 03 FEB 92 / UK Company News: Eurotunnel to seek damages for cost of extra safety

17: 178552: FT943-6917: FT 26 AUG 94 / Cross-Channel ferry blaze to be investigated
18: 1655: FT911-4602: FT 18 APR 91 / MMC to investigate Isle of Wight ferries
19: 5733: FT921-365: FT 30 MAR 92 / Hopes for ship data recorder
20: 26988: FT922-7334: FT 19 MAY 92 / World Trade News: Denmark-Sweden ferry link-up is agreed
21: 119826: FT933-1606: FT 23 SEP 93 / Ferry operator in link with Belgium
22: 150782: FT941-1125: FT 26 MAR 94 / International Company News: Vard plans to spin off ferry division
23: 118260: FT933-15867: FT 07 JUL 93 / New high-speed Stena ferry in service by 1995
24: 119845: FT933-1625: FT 23 SEP 93 / Sally Line agrees Belgian link-up
25: 32757: FT922-12800: FT 15 APR 92 / Freight ferry
26: 199245: FT944-10109: FT 12 NOV 94 / Tighter ferry rules proposed
27: 114042: FT933-11894: FT 27 JUL 93 / International Company News: Vard set to spin off ferry unit
28: 200184: FT944-11048: FT 08 NOV 94 / Bow doors faulty on 33% of ferries using UK ports: Government to increase safety checks on vessels
29: 62351: FT924-11264: FT 27 OCT 92 / Ferry operators accused of pricing collusion
30: 199989: FT944-10853: FT 09 NOV 94 / Eurotunnel hits at government on ferry safety
31: 143053: FT941-732: FT 29 MAR 94 / Netherlands ferry route may restart
32: 199238: FT944-10102: FT 12 NOV 94 / Tighter ferry rules proposed
33: 84325: FT931-16573: FT 06 JAN 93 / Cross-Channel ferries hint
34: 125074: FT934-497: FT 24 DEC 93 / International Company News: Greek ferry operator in cash call
35: 64622: FT924-13535: FT 15 OCT 92 / New ferry service
36: 28865: FT922-9211: FT 08 MAY 92 / New ferry is largest in Channel
37: 137406: FT934-1954: FT 16 DEC 93 / Technology: Ships bridge the danger gap - Andrew Fisher concludes a series on transport safety with an investigation into innovations that may help prevent sea disasters and give clues to their causes
38: 2421: FT911-5368: FT 15 APR 91 / World News in Brief: Ferries disrupted
39: 26728: FT922-7074: FT 20 MAY 92 / Boulogne freight link
40: 44107: FT923-9034: FT 07 AUG 92 / Ferry row settled
951 s, Reading doc 174533:FT943-3295, click from main win, time Tue Aug 12 09:40:48 1997
956 s, Reading doc 42744:FT923-7671, click from main win, time Tue Aug 12 09:40:53 1997
965 s, Reading doc 149044:FT941-12581, click from main win, time Tue Aug 12 09:41:02 1997
970 s, Reading doc 72637:FT931-5947, click from main win, time Tue Aug 12 09:41:07 1997
Doc number 72637 added to aspect 6
Aspect # 6, auto terms are congo brazzaville zairean “illegal immigrant“ “death toll“
U: drags “congo“ into label area
992 s, Reading doc 208393:FT944-18499, click from main win, time Tue Aug 12 09:41:29 1997
1006 s, Reading doc 208402:FT944-18508, click from main win, time Tue Aug 12 09:41:43 1997
1008 s, Reading doc 9804:FT921-686, click from main win, time Tue Aug 12 09:41:45 1997
1012 s, Reading doc 186180:FT943-1239, click from main win, time Tue Aug 12 09:41:49 1997
1016 s, Reading doc 207852:FT944-17958, click from main win, time Tue Aug 12 09:41:53 1997
1020 s, Reading doc 177939:FT943-316, click from main win, time Tue Aug 12 09:41:57 1997
1022 s, Reading doc 201958:FT944-12822, click from main win, time Tue Aug 12 09:41:59 1997
1031 s, Reading doc 208769:FT944-18875, click from main win, time Tue Aug 12 09:42:08 1997
1034 s, Reading doc 14222:FT921-11074, click from main win, time Tue Aug 12 09:42:11 1997
1041 s, Reading doc 178552:FT943-6917, click from main win, time Tue Aug 12 09:42:18 1997
1048 s, Reading doc 1655:FT911-4602, click from main win, time Tue Aug 12 09:42:25 1997
1054 s, Reading doc 5733:FT921-365, click from main win, time Tue Aug 12 09:42:31 1997
1064 s, Reading doc 26988:FT922-7334, click from main win, time Tue Aug 12 09:42:41 1997
1070 s, Reading doc 119826:FT933-1606, click from main win, time Tue Aug 12 09:42:47 1997
1079 s, Reading doc 118260:FT933-15867, click from main win, time Tue Aug 12 09:42:56 1997
1082 s, Reading doc 119845:FT933-1625, click from main win, time Tue Aug 12 09:42:59 1997
1089 s, Reading doc 32757:FT922-12800, click from main win, time Tue Aug 12 09:43:06 1997
1092 s, Reading doc 199245:FT944-10109, click from main win, time Tue Aug 12 09:43:09 1997
1096 s, Reading doc 114042:FT933-11894, click from main win, time Tue Aug 12 09:43:13 1997
1105 s, Reading doc 125074:FT934-497, click from main win, time Tue Aug 12 09:43:22 1997
U: does extensive interactions with 3-D window
19:06 F: *Could you say what you're doing there? With the 3-D window?*
U: *I'm just looking at it and trying to see how the articles I've picked lay out in this 3-D network. I'm just trying to figure out how I could make it more*

useful for my searching purposes. I'm really thinking about things how if I'm searching for things on the Internet and I had something like this how would I be able to use it. It's an interesting idea.
20:00 F: *Time's up.*
7 documents saved in 7 aspects, with 0 miscellaneous docs
Aspect 0, 1 docs saved
Auto Terms: estonia estonia_sink forsberg estonia_disaster bow_section
User supplied text = estonia
FT943-312: 177935 FT 30 SEP 94 / Ferries in six 'near accidents': Finland and Sweden order checks after Estonia sinking
Aspect 1, 1 docs saved
Auto Terms: ferry_sink ferry_disaster wedding_party high_sea bangladesh
User supplied text = bangladesh
FT944-15661: 205199 FT 17 OCT 94 / World News in Brief: Bangladesh ferry sinks
Aspect 2, 1 docs saved
Auto Terms: ferry_sink cargo_ship manila sink survivor
User supplied text = manila
FT944-5773: 194241 FT 02 DEC 94 / World News in Brief: Manila ferry sinks
Aspect 3, 1 docs saved
Auto Terms: bow_door marine_safety_agent ferry safety_agent dr_mawhinney
User supplied text = Herald of Free Enterprise
FT944-11048: 200184 FT 08 NOV 94 / Bow doors faulty on 33% of ferries using UK ports: Government to increase safety checks on vessels
Aspect 4, 1 docs saved
Auto Terms: ferry_sink port-au-prince neptune haiti product_centre
User supplied text = neptune
FT931-8485: 75524 FT 19 FEB 93 / Crowded ferry sinks off Haiti
Aspect 5, 1 docs saved
Auto Terms: moby imo vessel livorno ship
User supplied text = livorno
FT934-1954: 137406 FT 16 DEC 93 / Technology: Ships bridge the danger gap - Andrew Fisher concludes a series on transport safety with an investigation into innovations that may help prevent sea disasters and give clues to their causes
Aspect 6, 1 docs saved
Auto Terms: congo brazzaville zairean illegal_immigrant death_toll
User supplied text = congo
FT931-5947: 72637 FT 03 MAR 93 / World News in Brief: Congo ferry toll rises to 146
1200 s Tue Aug 12 09:45:21 1997
Stats from this run: 3 queries run
100 docs returned, 66 unique, 52 viewed
7 docs saved (including misc), 7 saved
saved docs:
FT931-5947: 72637 979
FT931-8485: 75524 466
FT934-1954: 137406 843
FT943-312: 177935 75
FT944-5773: 194241 255
FT944-11048: 200184 311
FT944-15661: 205199 162
saved good docs
FT931-5947: 72637 979
FT931-8485: 75524 466
FT934-1954: 137406 843
FT943-312: 177935 75
FT944-5773: 194241 255
FT944-11048: 200184 311
FT944-15661: 205199 162
Sparse Trec Data Starts HERE

1 FT943-312
2 FT944-15661
3 FT944-5773
4 FT944-11048
5 FT931-8485
6 FT934-1954
7 FT931-5947