Xerox Site Report: Four TREC-4 Tracks *

Marti Hearst, **Jan Pedersen**, **Peter Pirolli** and **Hinrich Schütze**
Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
{hearst,pedersen,pirolli,schuetze}@parc.xerox.com

Gregory Grefenstette and **David Hull**
Rank Xerox Research Centre
6, chemin de Maupertuis, 38240 Meylan, France
{gregory.grefenstette,david.hull}@xerox.fr

1 Overview

The Xerox research centers participated in four TREC-4 activities: the routing task, the filtering track, the Spanish track, and the interactive track. We addressed the core routing task as a problem in statistical classification: given a training set of judged documents, build an error-minimizing statistical classifier to assess the relevance of new test documents. This year, we built on the methodology developed in [21] by adding a combination strategy that pooled evidence across a number of separately trained classification schemes. Since many of our classifiers infer probability of relevance, adapting our routing methods to the filtering track consisted of obtaining probability estimates for the remaining classifiers and reporting those documents scoring above the probability thresholds determined by the three set linear utility functions.

Our contribution to the Spanish track focussed on the effect of principled language analysis on a baseline retrieval system. We employed finite-state morphology [14] and hidden-Markov-model-based part-of-speech tagging [7] to analyze Spanish language text into canonical stemmed forms, and to identify verbs and noun phrases. Various combinations of these were then fed into SMART [1] for ranked retrieval.

This year our activity on the ad hoc task focussed on the interactive track, which allows arbitrary user interaction in the process of finding relevant documents. We developed a graphical user interface to two interactive tools, Scatter/Gather [6] and Tilebars [11], and asked a number of subjects to use this tool to "find as many good documents as you can for a topic, in around 30 minutes, without collecting too much rubbish." We set up an experimental design to measure the value of each tool, and their combination, averaging out subject effects. That is, we were interested in determining how well the average user might perform with interactive tools rather than measuring the very best performance possible assuming an expert searcher.

These efforts are described in more detail in the following sections.

2 The Routing Problem

The routing task can be treated as a problem of machine learning or statistical classification. A classification tool is inferred from the training set of judged documents and is used to predict the relevance of newly arriving documents. Traditional learning algorithms must be adapted due to the large scale of this problem. For example, one cannot use the full collection vocabulary as a feature set without overfitting to the training documents. Similarly, the full training set is simply too large, and relevant documents too rare to efficiently learn the optimal classification rule. Therefore, Xerox has developed a special three-step algorithm to solve the routing problem (described in detail in [21]). In this section, we summarize the Xerox routing strategy and present our new work on method combination developed for TREC-4.

2.1 Step 1: Local Regions

The document collection is parsed, tokenized, stemmed, and stop words are removed using the Text Database System (TDB) developed at Xerox PARC [8]. Indexed *terms* consist of single words and two-word phrases that occur over five times in the corpus (where a phrase is defined as an adjacent word pair, not including stop words). This process produces over 2.5 million *terms*. Then, each document is partitioned into overlapping sections of average length 300 words with an average overlap of approximately 250 words.

*Authors listed in alphabetic order. Hull, Pedersen and Schütze worked on the routing task, Hull on the filtering track, Grefenstette and Hull on the Spanish track, and Hearst, Pedersen, and Pirolli on the interactive track.

In the first stage, expanded queries are constructed using a modified version of the Rocchio technique for relevance feedback [3]. The expanded query is defined as the vector sum of the relevant documents in the training set. The original query is also included and given a weight of five relevant documents. No negative feedback is used, i.e. non-relevant documents are ignored. All documents in the collection are ranked according to their similarity to the expanded query, and the top 2000 documents are selected for each query. These documents define the *local region* for that query. Documents are given the rank of their highest scoring section and for all future analysis each document in the local region is represented only by this section.

There are a number of advantages to using only the local region in subsequent stages of the routing task. First, the size of the training set is substantially reduced, so it is possible to solve the problem using computationally intensive variable selection techniques and learning algorithms in a reasonable length of time. Second, the density of relevant documents is much higher in the local region than in the training collection as a whole, which should improve classifier performance. Third, the non-relevant documents selected for training are those which are most difficult to distinguish from the relevant documents. These non-relevant documents are clearly among the most valuable ones to use as training data for a learning algorithm.

2.2 Step 2: Document Representations

In the vector space model, one dimension is reserved for each unique term in the collection. Standard classification techniques cannot operate in such a high dimensional space, due to insufficient training data and computational restrictions. Therefore, some form of dimensionality reduction must be considered, even after applying the preliminary filtering step to construct the local region. We use two distinctive approaches to dimensionality reduction, optimal term selection and reparameterization of the document space.

The process of optimal term selection consists of identifying the words that are most closely associated with relevance. Our approach is to apply a χ^2 test to the contingency table containing the number of relevant and non-relevant documents in which the term occurs (N_{r+} and N_{n+}, respectively), and the number of relevant and non-relevant documents in which the term doesn't occur (N_{r-} and N_{n-}, respectively).

$$\chi^2 = \frac{N(N_{r+}N_{n-} - N_{r-}N_{n+})^2}{(N_{r+}+N_{r-})(N_{n+}+N_{n-})(N_{r+}+N_{n+})(N_{r-}+N_{n-})}$$

The higher the score, the greater the association between that term and the relevant documents. We select the 200 terms with the highest scores for each query to use with our learning algorithms.

As an alternative approach, we apply Latent Semantic Indexing (LSI) [9] to represent documents by a low-dimensional linear combination of orthogonal indexing variables. The LSI solution is computed separately for each query by applying a singular value decomposition to the sparse document by term matrix constructed from the local region. We then select the 100 most important LSI factors to serve as an alternative document representation. Optimal term selection works best when there is a relatively small specific vocabulary associated with the topic while LSI performs well when the topic has a very large vocabulary which can be organized into a smaller number of general concepts.

2.3 Step 3: Classification Algorithms

We have examined a number of classification techniques for the routing problem, including logistic regression, nearest neighbors, linear discriminant analysis (LDA), and neural networks. In the past, we have achieved the best performance with LDA (using LSI features) and a linear neural network (using both LSI and term features). For TREC-4, we decided to try to improve performance by combining the results of a number of different classification techniques. We submitted two runs, a baseline using LDA and a combination run using using Rocchio expansion, nearest neighbors, LDA, and a linear neural network fitting a logistic model.

We briefly describe the three classification techniques. Our variant of Rocchio expansion is detailed in step 1 above. Linear Discriminant Analysis finds a linear combination a of the feature elements which maximizes the separation between the centroids of the relevant $\bar{x}_r$ and non-relevant $\bar{x}_n$ documents: $a = S^{-1}(\bar{x}_r - \bar{x}_n)$, where S is the pooled within-group covariance matrix. The linear neural network (no hidden units) uses backpropagation to iteratively fit a logistic model. The network is not iterated to convergence, rather a validation set is used to determine the optimal number of training iterations, as a protection against overfitting. Our nearest neighbor technique is a slightly modified version of Yang's Expert Network [22]. Essentially, each new document is assigned a score equal to the sum of its similarity scores with respect to all relevant documents among its 50 nearest neighbors in the training set.

In order to combine results, the output of the various classification techniques needs to be normalized to the same scale. We accomplish this by converting the output for each classifier into an estimated probability of relevance. The probability of relevance of each document can be extracted automatically from the model for LDA and the network fitting a logistic model. For the nearest neighbor technique and Rocchio expansion, we used logistic regression to transform the output values into probability estimates [16], i.e. $p(s) = \frac{exp(a+bs)}{1+exp(a+bs)}$ where a and

b are the parameter estimates obtained from the training data.

We examined three possible approaches to combining evidence.

(1) Averaging the probability estimates.

(2) Using linear regression on the training data to determine the optimal linear combination of the probability estimates on a per-query basis.

(3) As (2), but compute the average optimal combination and use it for all queries.

In preliminary experiments, the simple average of the probability estimates (1) worked best, and we used this approach in our submitted run.

2.4 Summary of Routing Algorithm

We present a simple flow chart that describes the training and testing process of our routing algorithm. The training process:

- Compute 2000 nearest documents to Rocchio expanded query (local region).
- Segment documents and select segment most similar to expanded query (on a per-query basis).
- Compute LSI decomposition of local region and select 100 largest factors.
- Compute χ^2 statistic to select 200 most valuable terms in local region.
- Apply classification techniques to documents in local region of training set using the appropriate representation.
- Convert the document scores to probability estimates.
- Average the probability estimates for the combination run.

The testing process:

- Select test documents whose best segment exceeds a threshold similarity score.
- Obtain LSI document vectors for selected test documents from LSI term representation.
- Score selected test documents using the classifiers and convert the scores to probability estimates.
- Average the probability estimates for the combination run.
- Rank the test documents by descending probability score. All documents in the selected region are ranked ahead of those that fall below the threshold similarity score.

2.5 TREC-4 Results

The basic Xerox TREC-4 routing results are presented in Table 1. The first two rows measure uninterpolated average precision at all relevant documents (all) while the second pair measure average precision at 20 documents retrieved (20). The column *thr* gives the threshold used to select documents for classification. For example, a threshold of 750 means that all documents which have a similarity score greater than the 750th training document (wrt the Rocchio expanded query) are selected. The other columns are: Roc = Rocchio expansion, NN = nearest neighbors, LDA = linear discriminant analysis, Net = logistic neural network, and Cmb = combination run. The submitted runs are marked in the table. The Cmb run submitted to NIST was partially corrupted due to a programming error. The corrected results are presented in the table.

	thr	Roc	NN	LDA	Net	Cmb
all	750	0.355	0.371	(0.384)	0.400	[0.397]
	2000	0.355	0.370	0.383	0.412	0.420
20	750	0.507	0.537	(0.562)	0.597	[0.595]
	2000	0.507	0.546	0.568	0.620	0.629

Table 1: Average precision at all relevant docs (all) and average precision at 20 docs (20) for various routing strategies. () - submitted xerox1, [] - corrected xerox2

As an alternative method of evaluation, we can rank the methods by their performance for each query. When these ranks are averaged, we get a measure that is less sensitive to extremely variable queries [1]. Table 2 presents the average rank for each method. The final column gives the rank difference that is statistically significant with a p-value of 0.05 according to the Friedman Test [13]. The test results should be taken with a grain of salt, since the Friedman Test assumes independence between methods (clearly violated here, since one method is a combination of the others!). However, it seems quite clear that the neural network and the combination run significantly outperform the other methods.

	thr	Roc	NN	LDA	Net	Cmb	sig.
all	750	2.08	2.76	2.74	3.58	3.84	0.51
	2000	2.19	2.59	2.68	3.48	4.06	0.53
20	750	2.40	2.61	2.86	3.58	3.55	0.45
	2000	2.27	2.64	2.76	3.60	3.73	0.48

Table 2: Average within-query rank of routing strategies.

It is valuable to look more closely at the choice of threshold. Previous research had found that average performance for individual classifiers was optimized by selecting a very restrictive threshold (750). The TREC-4

[1] See Hull [13] for reasons why this might be advisable.

results suggest that this is not always the case. Table 3 presents the fraction of queries which score better at a threshold of 2000 than at a threshold of 750 (queries with equal performance are ignored). For the individual classifiers, the less restrictive threshold hurts about as many (or more) queries as (than) it helps. However, the combination run is much more robust, allowing the system to apply advanced classification techniques to a much larger number of test documents, resulting in a corresponding improvement in performance.

	NN	LDA	Net	Cmb
all	0.383	0.425	0.511	0.652
20	0.500	0.458	0.588	0.733

Table 3: Percentage of queries where threshold 2000 is better than 750 (queries with equal performance are ignored).

Table 4 compares the performance of the Xerox routing system to other systems at TREC-4. In particular, it measures the absolute difference in uninterpolated average precision at all relevant documents between Xerox runs and the best result submitted to TREC-4 for each query. The submitted LDA run was within striking distance (5%) of the best performing system for 14 of the 50 queries and reasonably close (within 10%) for 27 of 50 queries. The best posthoc run (combination with threshold 2000) improves those numbers to 21 and 33 queries respectively. The system performs quite solidly but there is certainly room for improvement.

Another aspect that we would like to evaluate more closely is the effect of bias in the test set. The test set for TREC-4 consisted of the Ziff data from disk3 plus new Federal Registry documents. Since all of disk3 was part of the training set, this means that some queries were trained on data that was subsequently used to test performance. In particular, 6 out of the 25 Computer queries and 13 out of the 25 Federal registry queries had disk3 Ziff documents in the local region, and hence offer training sets potentially biased towards higher performance from our learning-intensive classifiers. Initial indications suggest that this is indeed the case for the biased Computer queries, but not the case for the Federal Registry queries.

	thr	$< 5\%$	$5-10\%$	$10-20\%$	$> 20\%$
LDA	750	14	13	15	8
Net	750	18	12	14	6
Cmb	2000	21	12	13	4

Table 4: Comparative results: absolute difference in uninterpolated average precision at all relevant documents (all) between method and best result submitted to TREC-4 for that query. Table values are number of queries within given range.

3 The Filtering Task

The filtering task is closely related to the routing task described in the previous section. The primary difference is that the system must also make a binary decision about whether to accept or reject each test document. Also, evaluation is based on a utility function of the form A*R-B*N where R and N are the number of relevant and non-relevant documents in the retrieved set and A and B are positive constants. The goal is to maximize this utility function. Three separate runs were submitted with utility functions (A,B) = (1,3), (1,1), and (3,1), which are optimized by selecting documents with an expected probability of relevance greater than p = .75, .5, and .25, respectively [17].

We estimate probabilities explicitly in the routing task as part of our combination strategy, as described in step 3 of the previous section. Therefore, to produce filtering results, we need merely filter the returned set according to these probability estimates. Note that only test documents which pass our initial thresholding step will be considered. In preliminary experiments, we found that a combination of the four classification techniques used in routing also worked well for filtering. However, we also discovered that the combination run tended to underestimate the actual probability of relevance, often quite substantially, and we obtained better performance by averaging the two largest probability scores for each document. Unfortunately, our filtering run submitted to NIST was also corrupted, since it was based on the same data used for the combination routing run. The results below have been generated from the corrected data.

P	thr	Roc	NN	LDA	Net	Cmb	Top2
75		2.64	3.30	3.66	4.20	3.43	3.77
50	750	2.50	3.00	3.47	3.75	4.07	4.21
25		2.61	2.87	3.52	3.62	4.17	4.21
75		2.84	3.33	2.87	4.31	3.69	3.96
50	2000	2.89	3.07	2.88	3.93	4.23	4.00
25		2.62	3.04	3.39	3.69	4.15	4.11

Table 5: Average within-query rank of filtering strategies.

Since the scale of the utility scores differs across queries, it is misleading to summarize the results simply by looking at the average utility. Instead, we compare classifiers using the average rank statistic, as presented in the previous section. The results are shown in Table 5. The new column Top2 is derived by taking the average of the largest two probability estimates for each document, and is designed to correct for the bias mentioned above. In general, the network and combination strategies tend to perform better than the alternatives, just as they did in routing. However, the combination run is much less effective (and the network run much more so) when the filter probability p = 0.75.

A filtering system must be capable of both ranking the documents accurately (the routing task) and selecting the proper size of the retrieved set. In order to compare between the routing and the filtering task, we attempt to separate performance due to ranking from performance due to selecting the right threshold. In order to measure this distinction, we take the ranked list returned by each classifier and compute the maximum utility which can be obtained from that ranking. Table 6 presents the average ranks for these optimal utility scores. By comparing the two tables, we learn that the combination run could score a lot better if it had more accurate probability estimates.

P	thr	Roc	NN	LDA	Net	Cmb	Top2
75		2.73	3.18	3.47	3.91	3.95	3.76
50	750	2.45	3.32	3.28	4.15	4.13	3.67
25		2.38	3.25	3.21	4.00	4.18	3.98
75		2.69	3.16	3.07	3.89	4.34	3.85
50	2000	2.48	2.74	3.09	4.10	4.54	4.05
25		2.25	2.71	3.02	3.97	4.79	4.26

Table 6: Average within-query rank of optimal filtering performance for each strategy.

In our preliminary experiments, we found that the probability estimates of the combination run often substantially underestimated the true probability of relevance. We obtained this information by comparing the observed number of documents in the relevant set to expected number, obtained by taking the sum of the probability estimates of the individual documents. Unfortunately, this method does not produce accurate summary statistics for the entire query set, because it does not include the queries where no documents are returned. An empty retrieved set often indicates that the system has underestimated the probability of relevance, which means that ignoring these queries may bias the average results. Therefore, we adopted a different approach. We computed the size of the optimal returned set for each query and method and compared it to the size of the actual returned set. Table 7 measures the percentage of the queries where the actual retrieved set is larger than the optimal retrieved set (ignoring queries where they are equal). Values much lower than 50% indicate that the technique is underestimating the probability of relevance for a large proportion of the queries and retrieving too few documents.

Table 7 reveals a number of interesting patterns. First, their is an overall tendency to return too few documents. Using the restrictive initial threshold of 750 is a large part of the problem. In hindsight, it is certainly a mistake to use a two-stage filtering algorithm and then expect the probability estimates used to filter in the second stage to be unbiased! The problem is most evident for thr = 750 and p=0.25, when many documents which are rejected in the first filtering stage have p>0.25 of being relevant.

P	thr	Roc	NN	LDA	Net	Cmb	Top2
75		0.418	0.316	0.591	0.290	0.186	0.318
50	750	0.396	0.326	0.532	0.370	0.302	0.500
25		0.245	0.245	0.327	0.367	0.327	0.449
75		0.418	0.461	0.682	0.333	0.186	0.452
50	2000	0.396	0.467	0.633	0.479	0.364	0.583
25		0.265	0.449	0.583	0.500	0.460	0.540

Table 7: Percentage of queries where actual retrieved set is larger than the optimal retrieved set (queries where they are equal are ignored).

When the initial threshold is increased to 2000, this bias is substantially reduced. It is also interesting to note the LDA behaves much differently from the other classifiers, tending to overestimate the size of the retrieved set.

We have already mentioned that the combination run performs much worse than expected for p = 0.75. From Table 7, it is very clear why. It is retrieving too few documents for over 80% of the queries. The problem comes from the fact that the combination run is constructed by taking the average of the probability estimates from four different classifiers, which means that its variance will be four times smaller. Since documents with p>0.75 are extreme in any event, far fewer documents are likely to satisfy this criterion for the combination run. Therefore, we can conclude that using the averaged probability estimates directly is not the right approach for filtering. Instead, we need to renormalize the estimates, perhaps by using logistic regression over the training set, in order to rescale the variance. Our naive initial attempt to correct for underestimation by using the top two probability estimates for each document seems to work reasonably well, but the optimal performance table tells us that the ranking produced by this measure is poorer, so we will be better off in the long run if we correct the probability estimates obtained for the original combination strategy.

4 Spanish Track

Our approach to Spanish is traditional from an information retrieval perspective. We are interested in testing whether our Spanish linguistic tools improve retrieval performance. For this work, we use SMART [1] as the underlying text retrieval system, but with all our linguistic analysis being done prior to indexing and retrieving documents. New fields are created for each document containing the different linguistic components we derive from the original text.

We use the following linguistic tools, developed at Xerox and Rank Xerox, to analyze the Spanish text:

- a morphological analyzer [15], which returns part of speech information and lemmatized forms of individual words (e.g. países → país +Noun+Masc+Pl)

```
<DOC> <DOCNO> SP94-0000662 </DOCNO>
<ARTNUM> 0000662 </ARTNUM>
<HEADLINE> San Marcos espera de NL arte joven          </HEADLINE>
<TEXT> La Subsecretaria de Cultura reune obras de artistas locales
       menores de 30 anos para enviarlas al certamen de la feria de
       Aguascalientes.                                 </TEXT>
<F1>   el subsecretaria de cultura reunir obra de artista local menor
       de 30 ano para enviar al certamen de el feria de Aguascalientes .
                                                       </F1>
<F2>   obra de artista local menor
       ano
       certamen
       feria de Aguascalientes                         </F2>
<F3>   reunir
       enviar                                          </F3>
<F4>   subsecretaria_de_cultura
       obra_de_artista_local_menor
       ano
       certamen
       feria_de_Aguascalientes                         </F4>  </DOC>
```

Figure 1: Sample Spanish document augmented with three additional fields: lemmatized words, noun phrases, lemmatized verbs, joined noun phrases. This format allows us to test different combinations of linguistically-derived information, without reindexing the corpus.

- a part-of-speech tagger [7], which uses the morphological analyzer and a trained hidden Markov model (HMM) network to choose the part-of-speech of a word from context
- a noun phrase extractor, which extracts noun phrase patterns from tagged text.

The major problem that we face in our Spanish experiments is dealing with accented text: some words that should have been accented were not, and we have had trouble getting SMART to recognize some accented characters. To solve the first problem, we modified our linguistic tools so that they correctly lemmatize unaccented words. For the second, we strip off accents before feeding the text to SMART. Fortunately, there are very few confusions between accented and unaccented words in Spanish.

Our treatment of a Spanish document begins by part of speech tagging the TEXT field contents. Each tagged word is then lemmatized according to its part of speech. From this tagged text, a number of supplementary fields are created and added to the original document: field F1 contains a lemmatized form of the text of field TEXT, field F2 contains all the lemmatized noun phrases of one word or more (one per line), field F3 contains all and only the lemmatized verbs, and field F4 contains all the lemmatized noun phrases with intervening spaces replaced by underscores so that they will be considered as units by SMART. Figure 1 gives a sample document. The total time needed to create these augmented documents over the 68,000 Spanish documents (200 MBytes) is 39 hours of real time on a SPARC 20. We treat the queries in a similar fashion.

We then conducted a preliminary analysis on the first 25 Spanish queries to determine which approaches were most successful. We submitted two runs, a baseline and one constructed using query expansion. The baseline runs uses the contents of fields F1 and F2, which corresponds to using lemmatized text and doubling the weight of the component terms in noun phrases. Indexing noun phrases by their components, rather than treating them as single units, produced slightly superior performance in the preliminary tests, which motivates this decision. Unfortunately, a programming error resulted in parts of some noun phrases being truncated from field F2, so the submitted run does not precisely match the desired experiment.

Since the the new Spanish queries are particularly short, there is reason to hope that they might benefit from some query expansion. Our approach is to take the first 20 documents returned by the baseline run and extract all terms that occur significantly more often in this document sample than one would expect by chance. The terms are selected according to a binomial likelihood ratio test [10], comparing their occurrence in the first 20 documents to their occurrence in the rest of the collection. The selected terms are then weighted in proportion to the significance of their occurrence in the sampled documents. Since it uses the baseline results, this run may also be affected by the programming error described above.

	query set	base	infl	infl-np	expand
all	Q1-25	0.454	0.484	0.492	0.467
	Q26-50	0.174	0.204	0.212	0.267
20	Q1-25	0.718	0.718	0.722	0.722
	Q26-50	0.306	0.354	0.378	0.402

Table 8: Average precision at all relevant docs (all) and average precision at 20 docs (20) for Spanish queries.

The corrected Spanish performance figures are presented in Table 8. We include four different runs: (1) base = stop list but no morphological analysis, (2) infl = text lemmatized (stemmed) with inflectional morphology, (3) infl-np = noun phrase weight doubled, and (4) expand = query expansion. The uncorrected performance figures for infl-np and expand on Q26-50 (corresponding to our submitted runs) are 0.190/0.366 and 0.238/0.380 respectively. We present separate results for queries 1-25 (used for TREC-3) and 26-50 (used for TREC-4) since the former are substantially longer, and we note that the results reflect the difference in length.

We find that lemmatization using inflectional morphology helps in most cases, making a 3-5% absolute difference in performance. However, when the queries are long and the user is examining fewer than 20 documents, there is no improvement. These conclusions agree with the results obtained for English [13], although the Spanish inflectional morphology is somewhat more effective than its English counterpart. Doubling the weight of noun phrases only slightly improves performance. Our query expansion technique is harmful for the long queries, but improves performance quite substantially for the short queries. Unfortunately, this improvement tends to be restricted to the queries where we are already doing well, so the value of this automatic expansion technique is limited.

When compared to other systems, the corrected infl-np run is more consistent, scoring above the median for all 25 queries, but always well below the best performing system. The corrected expansion run scores as well as the best system for 3 queries, but it is also below the median for 3 queries, as it tends to drag down performance for queries where no good expansion terms can be found. This suggests that we should look for an expansion technique that provides more consistent (if less dramatic) improvements in performance. In general, the Spanish linguistic tools provide solid though unspectacular benefits for the information retrieval problem.

5 Interactive Track

5.1 Goals of the Experiment

The interface used in this session represents the first time we have integrated Scatter/Gather [6] with TileBars [11] and Ranked Titles (standard similarity search via the vector space model). Users can display retrieval results in these three modes, each with a different ranking strategy, and the output of one mode can be used as input to another mode. The goal of the interface was to provide multiple ways for the users to view retrieval results, in the expectation that different modes and ranking orders are appropriate at different points in the search, and that the usefulness of a mode varies with the kind of query being investigated.

We predicted that subjects would use the clusters produced by Scatter/Gather to organize the initial retrieval results and select subsets of these results for further examination (or to eliminate subsets from consideration) and then use TileBars to help determine which documents are relevant to the query. We also suspected that users would make little use of the display of ranked titles given the TileBar visualization and ranking as an alternative.

We originally planned to run experiments that would test each interface mode individually, and to examine the effectiveness of each mode on particular queries, but this would have required more subject hours than could be accommodated in the time available.

5.2 The System

Our system consists of the TDB [8] (Text DataBase) system developed at PARC and a new user interface that combines standard vector space search, Scatter/Gather, and the TileBar display method. TDB is implemented in Common LISP and CLOS, and the interface is implemented in TCL/TK [19]. The two parts communicate with one another through ILU [5] and expectk [18].

A flow diagram of the process model for the Interactive Track Interface is shown in Figure 2. First the user specifies a query, (in the form of a list of topics, see below). A threshold k is set indicating how many documents are to be retrieved initially. Then the k top-ranked documents, according to the vector space model, are retrieved and shown to the user in Title Mode. After this, the user can switch the display of the retrieval results among the three modes of Titles, TileBars, and Scatter/Gather. The user can view a subset of the retrieval results by selecting one or more of the clusters produced by Scatter/Gather, thus indicating that only the contents of those clusters are to be viewed. (The system keeps track of state information and allows the user to back up to previous states.) The user can reformulate the description of the query if desired. At all points, document titles can be marked as relevant. At the end of the session, the documents so marked are saved into a file.

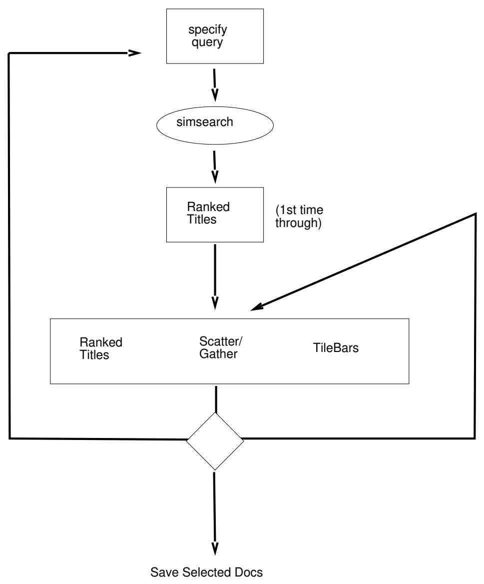

Figure 2: A flow diagram of the process model for the Interactive Track Interface.

As stated above, TDB provides a standard vector space weighting and ranking scheme, similar to that reported in [2] as well as standard Boolean search. It also includes support for Scatter/Gather clustering, as well as something we call structured similarity search, which provides support for TileBars. Structured similarity search works the same way as standard similarity search (the vector space model), except that it returns a list of term offsets that correspond to the query term sets as described below.

Our system as shown to users did not include relevance feedback or query expansion via suggestions of related terms although these mechanisms fit within our framework and will be included in future. Relevance feedback would most likely have improved the results, but we had not yet incorporated this mechanism with the structured similarity search at the time the experiments were run.

5.3 The User Interface

Our interface focuses on helping users understand and explore their retrieval results, rather than concentrating on query formulation and refinement. This is not to say that query reformulation is unimportant, but rather that we chose not to emphasize that for these experiments. Furthermore, we have evidence that the Scatter/Gather interface helps users determine alternative terms by which to augment their queries, see [20].

Figure 3 shows the entire interface when set in TileBar mode. We will use query 216 as a running example with which to illustrate the components of the interface:

> *What research is ongoing to reduce the effects of osteoporosis in existing patients as well as prevent the disease occurring in those unafflicted at this time?*

The lefthand side of the interface contains the user's query specification, the system's translation of the query, an active log of the state of the system as the user switches from mode to mode, and a window showing the documents the user has saved for the query. There is also a radiobutton that allows the user to specify the document cutoff threshold for the vector space search. The righthand side of the interface consists of a set of buttons that allow the user to change the mode of the retrieval results display, the display of the retrieval results, and a window for displaying document contents.

In Figure 3 the user has selected two documents as relevant, as indicated by the dark circles, and their titles appear in the "Saved Relevant Documents" window.

5.3.1 Specifying the Query

To accommodate the TileBar interface, users are required to enter their queries into a sequence of entry windows, as shown in the upper lefthand corner of the interface. The user is told that each line should correspond to a different topic of the query. Each entry line is called a "termset" since it is meant to contain a set of terms representing one topic. In this example, the participant has broken query 216 into four topics (in this case there is only one word per topic): *osteoperosis, patient, prevention,* and *research.*

In other recent work [12], we have found that we can achieve very strong improvements in precision at high document cutoff levels by first requiring the query to be specified in terms of a list of topics, and then applying the following two constraints:

- Treat the list of topics as a Boolean conjunct of disjuncts,
- Apply a proximity constraint of 100 tokens to this conjunct of disjuncts

In other words, the terms in each topic are treated as a disjunction, and a conjunction is imposed among the topical disjuncts. Only the documents which pass through the filter of these two constraints are retained, and they are ranked according to the vector space ranking, as before. The topmost termset is considered most important, and termsets lower down in the list are considered less important (some variations of this algorithm in which not all termsets are required have been experimented with). This algorithm yielded very strong improvements over using the short query with no constaints, or on using the original query specification, in the case of TREC 4 queries.

Similar results were found independently by the Waterloo group at TREC 4 [4]. In that work, the queries were again specified as an ordered list of topics, called subqueries. Most subqueries were specified as disjuncts. Each subquery was ranked separately, and then the ranks were combined by using the ranking of the topmost first, the secondmost second, and so on. The proximity constraint was slightly different in this algorithm – instead of being fixed, the documents were ranked according to the inverse of the distance between pairs of subqueries. The results obtained for the manual ad hoc using this approach were quite strong.

In this set of experiments, however, we did not treat the query as a Boolean query, but rather converted it into a bag of words and performed a standard similarity search on the bag. The separate lines were made use of, however, in order to display the corresponding hits in the TileBar display, described below.

5.3.2 Retrieval Result Modes

Figure 4 shows the initial results of the query as displayed in Title Mode, shown in ranked order according to the vector space model. The rank of the document is shown along side its title. When the document had no preassigned title then its document id label is shown instead. In this example, many of the top-ranked titles are very uninformative.

Figure 5 shows the results of running Scatter / Gather on the 250 retrieved documents. The results are partitioned into four clusters, of the sizes shown. Associated with each cluster is a list of "topical terms," extracted from the cluster centroids. These terms are meant to indicate the central topics encompassed by the documents within the cluster. In the figure we can see that the topical terms of Cluster 1, e.g., *bone osteoporosis, fracture, estrogen*, indicate that it should contain relevant documents. The topical terms for Clusters 3 and 4 make them appear to contain non relevant documents, whereas Cluster 2 may contain some relevant documents as it discusses *aid, cancer, institute, heart, center*, etc.

When the user selects Cluster 1 and then switches to TileBar mode, the view of Figure 3 results. The subset of 19 documents from Cluster 1 appear in the TileBar display. The user can use the Backup button to go back to the previous state, as shown in the History window. The TileBar representation works as follows. Each rectangle represents a document. Each row of the rectangle represents the corresponding termset in the query display, i.e, the top row corresponds to *osteoporosis*, the second row to *patient*, etc. Each row of each rectangle is comprised of a sequence of squares. Each square indicates a segment of the document; the leftmost square indicates the first segment, or paragraph, or other unit, of the document, the square to the right of this indicates the second segment of the document, and so on. The darkness of the square corresponds to the number of times the query occurs in that segment of text; the darker the square the greater the number of hits. White indicates no hits on the query term. Thus the user can quickly see if some subset of the terms overlap in the same segment of the document.

The TileBar ranking order is different than that of the vector space method. In this implementation, documents are first ranked by number of segments in which hits for all termsets overlap, second by the total number of hits in the document, and third by the vector space ranking.

As can be seen in Figure 3, the terms of the query are highlighted in the document display window; each color corresponds to a different line of the user query. In Figure 3 the contents of document FR88513-0157 appear in the appropriate window. This document was ranked highest among the 19 documents according to the TileBar ranking strategy. The portion of the document shown is one where the term hits overlap. A sentence is visible that states "Research is revealing that prevention may be achieved through estrogen replacement therapy for older women ..." and the rest of the context indicates that thing to be prevented is osteoporosis. Unfortunately, the TREC judges did not mark this document as relevant, perhaps because the article also contains a discussion of "National Osteoporosis Prevention Week". However, the TileBars strongly indicate that research is discussed in the same context as osteopororis and prevention in this document; perhaps use of such a tool could aid in the relevance judging process.

5.3.3 Interface Design Issues

We chose to lay out the components of the interface in one large window, in an attempt to keep in view all pertinent information at all times. This is as opposed to making use of menus and pop-up windows. We also wanted keep the number of possible mouse operations small. This information-constancy effect, although difficult to achieve due to limited screen real estate, acted as a useful constraint on the design of the interface.

As an example of the usefulness of this constraint, we decided that we wanted a clear, simple way for the user to select relevant documents. We decided to place a "checkbox" next to each document rather than requiring the users to perform cut-and-paste operations, or requiring them to remember which mouse button corresponded to selecting a document as relevant (clicking on a document title brings up the document for display rather than selecting that document as relevant). Several somewhat unexpected benefits resulted from this design choice. First, a marked checkbox is very distinctive visually. Second, the marks give an impression of the relative positions of the relevant and nonrelevant documents when a subset is reranked as the result of switching display modes. Third, the salience of the marked boxes potentially helps the user maintain some understanding the state of the search – more dark spots indicating more success with the search. Finally, when used in conjunction with the cluster display, the marked boxes sometimes served to signal a cluster that is interesting by virtue of the fact that it contains many relevant documents. One participant in our study remarked: "Sometimes I'd just go through the titles and select the ones that obviously pertained to the query, and then I would look at the clusters and see if they were concentrated at all. And sometimes that would happen and sometimes not."

Our participant interviews indicated that some of the participants did not like this all-in-one layout. This opinion may be caused by factors not directly related to usability, e.g., it may have been voiced because the participants were used to software that makes use of small pop-up windows, and because the fonts were too small (a comment given by some of the participants).

5.4 Experimental Design

Good experimental design requires that many different participants for each (query, system) combination, in order to reduce the effects of individual variation. Unfortunately, because each query requires 30 minutes to run, and because the rules of the track require that all 25 queries be covered, time constraints limited the amount of replication possible.

We attempted to maximize the experimental validity of our results while at the same time meeting the requirement of having users complete all 25 TREC queries. Our study consisted of four UC Berkeley graduate students, each of whom executed 13 queries. These consisted of 12 of the required 25, as well as one extra query given to all four participants. Only two of the participants' results on this query were reported, chosen arbitrarily.

We evaluated the queries in advance in order to rate them according to expected difficulty in general and with each search mode type. This prediction was handicapped in that it made use of restricted earlier versions of the interfaces and was indexed over only a subset of the collection used in TREC 4. We used these predictions for the experimental design, in order to ensure that each participant received a mix of "interface favorable" and "interface unfavorable" queries. These classifications can also be read as "easy" vs. "difficult".

Given these constraints, we developed a nested-factorial design to maximize within-participant measurement. In this design we nested, or split, queries across pairs of participants. Queries were assigned to participants as shown in Table 9. Participants were exposed initially to easy queries in each session, and these were then followed by harder ones.

Participants completed queries in two sessions. The experiments were run in an otherwise empty room with a video camera recording the session. Participants were given an 10-minute demonstration of the interface followed by a 10-minute warmup exersize, and the participants were provided with a 3 page description of the interface for reference. Additionally, a binder of topic descriptions was prepared for each participant, with each topic description appeared on the top of a separate page. Participants were not allowed to look at a new topic before the current one was completed.

Only 30 minutes was allowed per query, as specified in the instructions for the TREC interactive track. Participants were allowed to take short breaks between queries if desired. The instructions for the task were given as in the interactive track specifications "find as many good documents as you can for a topic, in around 30 minutes, without collecting too much rubbish." We took this as a hard time limit, and participants were required to stop when the 30 minute time limit was up. This statement emphasizes the finding of many relevant documents and deemphasizes the undesirability of including nonrelevant documents, and this had ramifications for how the participants performed. Some participants saved large numbers of documents for some of the queries without checking carefully for relevance, thus lowering overall precision.

We logged a good portion of the participants' activity, including which search state and mode type (Scatter/Gather, Tilebars, Titles) was in use when an action took place. We chose not to record every mouse event but did record when the search window or the document viewer windows were scrolled, as well as when clusters were selected and when documents were saved (and unsaved) as relevant. Since we recorded the visible contents of the search windows, we can make inferences about what documents were in view and what actions the user took in response to this information.

5.5 Results

5.5.1 Precision and Recall

The per-query results for precision and recall are shown in Tables 10 and 11. Our participants performed strikingly well on those queries with relatively few possible relevant documents that we predicted the interface would be helpful with, e.g., 207 and 216, and did poorly in the converse, e.g. 243, 236, 232, and 208. The relation between predicted difficulty and actual performance are more difficult to interpret for some queries, such as 220 and 216, but our system appears to have done as well as or better than other systems on these queries. Overall, the predictions

	A			B			A	A			B		
S1	210	202	207	203	204	205	212	211	216	215	206	208	209
S2	213	220	227	223	232	236	212	214	242	250	238	239	243
S3	210	202	207	203	204	205	212	211	216	215	206	208	209
S4	213	220	227	223	232	236	212	214	242	250	238	239	243

Table 9: Experimental Design. Queries were classified in advance as "interface favorable" (A) or "interface unfavorable" (B) and each participant was given a mix of the two types in the order shown, left to right. Four participants were used and each searched on half of the required queries.

Topic	**Ret**	**Rel**	**RR**	**Prec.**	**Recall**
202	19	283	13	0.684	0.045
203	6	33	1	0.166	0.030
204	5	397	3	0.600	0.007
205	4	310	1	0.250	0.003
206	2	47	2	1.000	0.042
207	67	74	41	0.611	0.554
208	6	54	2	0.333	0.037
209	16	87	8	0.500	0.091
210	36	57	27	0.750	0.473
211	26	323	25	0.961	0.077
212	21	153	18	0.857	0.117
213	12	21	5	0.416	0.238
214	3	5	3	1.000	0.600
215	26	183	23	0.884	0.125
216	24	36	17	0.708	0.472
220	10	24	5	0.500	0.208
223	26	363	14	0.538	0.038
227	85	347	71	0.835	0.204
232	1	9	0	0.000	0.000
236	14	43	0	0.000	0.000
238	48	270	28	0.583	0.103
239	100	123	20	0.200	0.162
242	11	38	6	0.545	0.157
243	15	69	1	0.066	0.014
250	25	86	10	0.400	0.116
TOTS/	608	3435	344	0.5658/	0.1001/
AVGS				0.5357	0.1570

Table 10: Scores determined by NIST for run XERINT1. Note that this represents the results of two different participants, paired arbitrarily. Ret = number retrieved, Rel = number possible relevant, RR = number retrieved that were relevant, Prec. = precision. Both macro and micro averages are shown for precision and recall.

Topic	**Ret**	**Rel**	**RR**	**Prec.**	**Recall**
202	39	283	12	0.307	0.042
203	15	33	3	0.200	0.090
204	5	397	4	0.800	0.010
205	1	310	1	1.000	0.003
206	10	47	7	0.700	0.148
207	44	74	32	0.727	0.432
208	4	54	2	0.500	0.037
209	24	87	13	0.541	0.149
210	33	57	27	0.818	0.473
211	22	323	19	0.863	0.058
212	23	153	11	0.478	0.071
213	19	21	7	0.368	0.333
214	4	5	3	0.750	0.600
215	42	183	36	0.857	0.196
216	29	36	13	0.448	0.361
220	15	24	11	0.733	0.458
223	37	363	2	0.054	0.005
227	52	347	46	0.884	0.132
232	3	9	2	0.666	0.222
236	41	43	6	0.146	0.139
238	55	270	30	0.545	0.111
239	20	123	8	0.400	0.065
242	10	38	9	0.900	0.236
243	16	69	1	0.062	0.014
250	22	86	7	0.318	0.081
TOTS/	585	3435	312	0.5333/	0.0908/
AVGS	585	3435	312	0.5629	0.1791

Table 11: Scores determined by NIST for run XERINT2. Note that this represents the results of two different participants, paired arbitrarily. Ret = number retrieved, Rel = number possible relevant, RR = number retrieved that were relevant, Prec. = precision. Both macro and micro averages are shown for precision and recall.

about the easy vs. difficult queries were born out. The average precision and recall scores for A vs. B queries were as follows:

A1: Prec .65 Rec .25

A2: Prec .61 Rec .33

B1: Prec .38 Rec .04

B2: Prec .50 Rec .08

What remains to be determined is whether or not other systems found the same queries easy and difficult in order to help determine if this effect is a function of the query, the interface, or both.

5.5.2 Participant Interviews

After the sessions the participants were interviewed about the use of the interface, and the results of these interviews were recorded and transcribed. The answers to the questions were quite informative. In answer to "What did you like best about the interface?" the participants answered as follows.

All four participants said explicitly that they liked the TileBar Interface or found it to be useful. One said: "I really like the tilebars the best. That's something unique. You can just click on it and get to that part of the document, and that's nice. It's like a magnifying glass." Another participant, while finding the TileBars useful, pointed out a problem with them, that sometimes even if the terms overlap, they do not necessarily overlap in a useful way and the visualization does not distinguish these two cases.

One participant had an interesting comment to make about the format in which queries were entered, saying: "I think having the four term sets is very useful. It was limiting, but on the other hand it really makes you think of the most important terms."

One participant was especially enthusiastic toward the clustering, finding the clusters useful for weeding out non-relevant documents, but did express concern about tossing out appropriate documents. Two participants mentioned liking the "sticky" checkboxes for selected relevant documents even after multiple searches on the same query. One mentioned the usefulness of the multi-color term highlighting in the display documents, where each color corresponds to a different query termset.

The participants were also asked "Is there anything [else] you didn't like about the interface?" In answer to this question and some of the others, we learned that system performance was one of the biggest problems. All four participants said that if the search performance had been better they would have done more searches. Two participants thought the sizes of the TileBars and the titles should be larger. One wanted keyboard accelerators. One participant was frustrated by the lack of a search abort capability. Two participants wanted a NOT operator and a phrase specification facility. One other participant asked for explicit AND and OR operators. All four thought that a term suggestion facility might have helped, but to differing degrees.

When asked how and when they used the TileBar facility, the participants answered as follows (these answers reveal information about the search strategies in general):

> "Basically, I used it after I had narrowed down the search a bit with the clusters. I usually used it to select relevant articles. I found the tile conjunctions useful to find phrases like "rain forest" but they weren't perfect. The tiles weren't always helpful if you had a fairly common word, or if it could be used in another way."

> "Almost all the time. It is the quickest way to tell which are the documents that have the most key words in it. If you get bored and don't want to read anymore, that's the quickest way to go."

> "That was usually a final part of the task. I would usually do the keyword selection, then I would do a clustering either once or twice depending on the results. Then I would go to the tilebar last and sometimes I would go to the tilebar mode, but not use it. I would just start scanning the titles and occasionally look at the tilebars. Other times I would really heavily use the tilebars. It just depended on the nature of the search."

> " ... I usually used it at the end. So once I got down to under 30, or around 30, documents then I'd just go look at the tilebars. ... I was looking for the conjunction of rain and forest and climate. Because not very many articles had all three together. But usually I'd wait until I only had a small clustering."

When asked how and when they used the Scatter/Gather display, the participants said they mainly used them to narrow down the set of articles to be viewed with TileBars and to eliminate unpromising documents. Large clusters were often reclustered. None of the participants thought having more than five clusters would be a good idea. Some users interwove the use of Scatter/Gather and TileBars.

When asked how and when they used the title display, all four participants said they didn't use it much because the other methods were more descriptive. Additionally, two participants said that the rankings were unhelpful and often misleading.

5.6 Conclusions

We find these preliminary results to be very encouraging; the participants performed well compared to the initial results returned for the other systems, were able to learn to use the new interface modes with very little acquaintance, and were enthusiastic about the new modes. The results of the participant surveys, a portion of which is reported in the preceding section, have given us very useful feedback about the merits of the interface and how it can be improved. On the top of the list is to add a term expansion suggestion mechanism, relevance feedback, improve the representation of the cluster contents, and improve search time performance.

We are currently devising measures to assess the usefulness of the display modes in various situations, based on choices users made given how many relevant documents were in view. We would also like to have available detailed transcripts of users of other systems, in order to help understand which kinds of displays are most helpful. Finally, we may conduct experiments in which only one of the three modes is available to facilitate evaluation of the effectiveness of each mode type.

Acknowledgements

We would like to acknowledge the efforts of Christine Diehl who conducted the participant studies, helped create the materials that were shown to the participants, helped determine the query difficulty rankings, transcribed the subject interviews and gave helpful comments on the design of the interface. Patricia Schank, now of SRI, also helped with the design of the experiment, determination of query difficulty, and commented on the design of the interface.

References

[1] Chris Buckley. Implementation of the smart information retrieval system. Technical Report 85-686, Cornell University, 1985.

[2] Chris Buckley, James Allan, and Gerard Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In Donna Harman, editor, *Proceedings of the Second Text Retrieval Conference TREC-2*. National Institute of Standards and Technology Special Publication 500-215, 1994.

[3] Chris Buckley, Gerard Salton, and James Allan. The effect of adding relevance information in a relevance feedback environment. In *Proc. 17th Int'l Conference on R&D in IR (SIGIR)*, pages 292–300, 1994.

[4] Charles L. A. Clarke, Grodon V. Cormack, and Forbes J. Burkowski. Shortest substring ranking (multitext experiments for TREC-4. In Donna Harman, editor, *Proceedings of the Fourth Text Retrieval Conference TREC-4*. National Institute of Standards and Technology Special Publication, 1996. (this volume).

[5] A. Courtney, W. Janssen, D. Severson, M. Spreitzer, and F. Wymor. *Inter-Language Unification, release 1.5.* Xerox PARC, 1994. ftp://ftp.parc.xerox.com/pub/ilu/ilu.html.

[6] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. 15th Annual Int'l ACM SIGIR Conference on R&D in IR*, June 1992. Also available as Xerox PARC technical report SSL-92-02.

[7] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. *Proc. of the Third Conference on Applied Natural Language Processing*, April 1992.

[8] Douglass R. Cutting, Jan O. Pedersen, and Per-Kristian Halvorsen. An object-oriented architecture for text retrieval. In *Conference Proceedings of RIAO'91, Intelligent Text and Image Handling, Barcelona, Spain*, pages 285–298, April 1991. Also available as Xerox PARC technical report SSL-90-83.

[9] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[10] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[11] Marti A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO, May 1995. ACM.

[12] Marti A. Hearst. Improving full-text precision using simple query constraints. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, Las Vegas, NV, April 1996. (to appear).

[13] David Hull. Stemming algorithms - a case study for detailed evaluation. *Journal of the American Society for Information Science*, 1996. To appear in special issue on the evaluation of IR systems.

[14] Ronald M. Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, September 1994.

[15] Lauri Karttunen. Constructing lexical transducers. In *Proc. of the 15th International Conference on Computational Linguistics, COLING'94*, Kyoto, Japan, 1994.

[16] K.L. Kwok, L. Grunfeld, and D.D. Lewis. Trec-3 adhoc, routing retrieval and thresholding experiments using pircs. In *Proceedings of the 3nd Text Retrieval Conference (TREC-3)*, pages 247–255, 1995.

[17] David Lewis. Evaluating and optimizing autonomous text classification systems. In *Proc. 18th Annual Int'l ACM SIGIR Conference on R&D in IR*, pages 246–255, 1995.

[18] Don Libes. expect: Curing those uncontrollable fits of interaction. In *Proceedings of the Summer 1990 USENIX Conference*, Anaheim, CA, June 1990.

[19] John Ousterhout. An X11 toolkit based on the Tcl language. In *Proceedings of the Winter 1991 USENIX Conference*, pages 105–115, Dallas, TX, 1991.

[20] Peter Pirolli, Patricia Schank, Marti A. Hearst, and Christine Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, WA, May 1996. ACM. (to appear).

[21] Hinrich Schütze, David Hull, and Jan Pedersen. A comparison of document representations and classifiers for the routing problem. In *Proc. 18th Int'l Conference on R&D in IR (SIGIR)*, pages 229–237, 1995.

[22] Yiming Yang. Expert network: Effective and efficient learning form human decisions in text categorization and retrieval. In *Proc. 17th Int'l Conference on R&D in IR (SIGIR)*, pages 13–22, 1994.

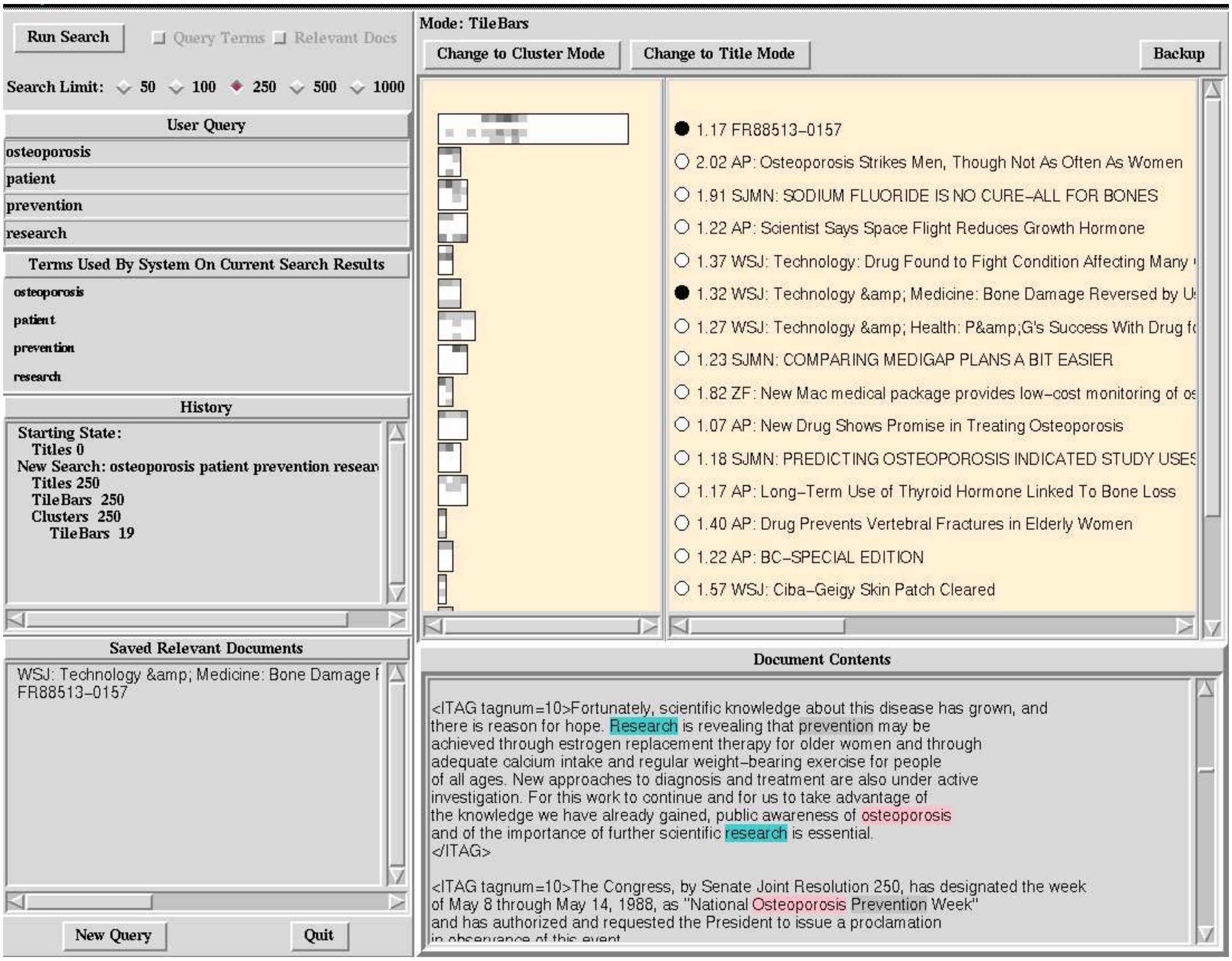

Figure 3: The PARC TREC 4 Interactive Interface, in TileBar mode.

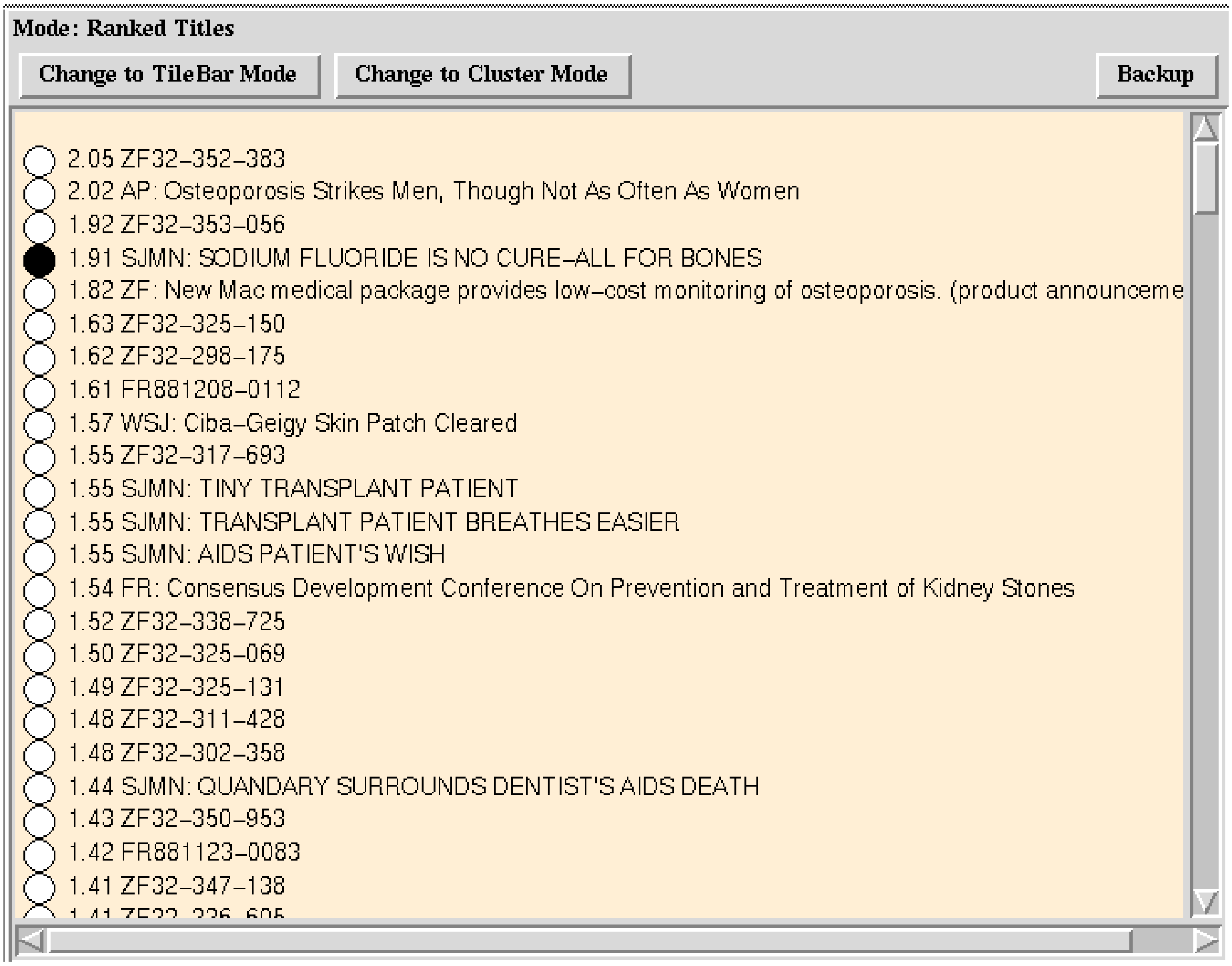

Figure 4: The interface in Title mode.

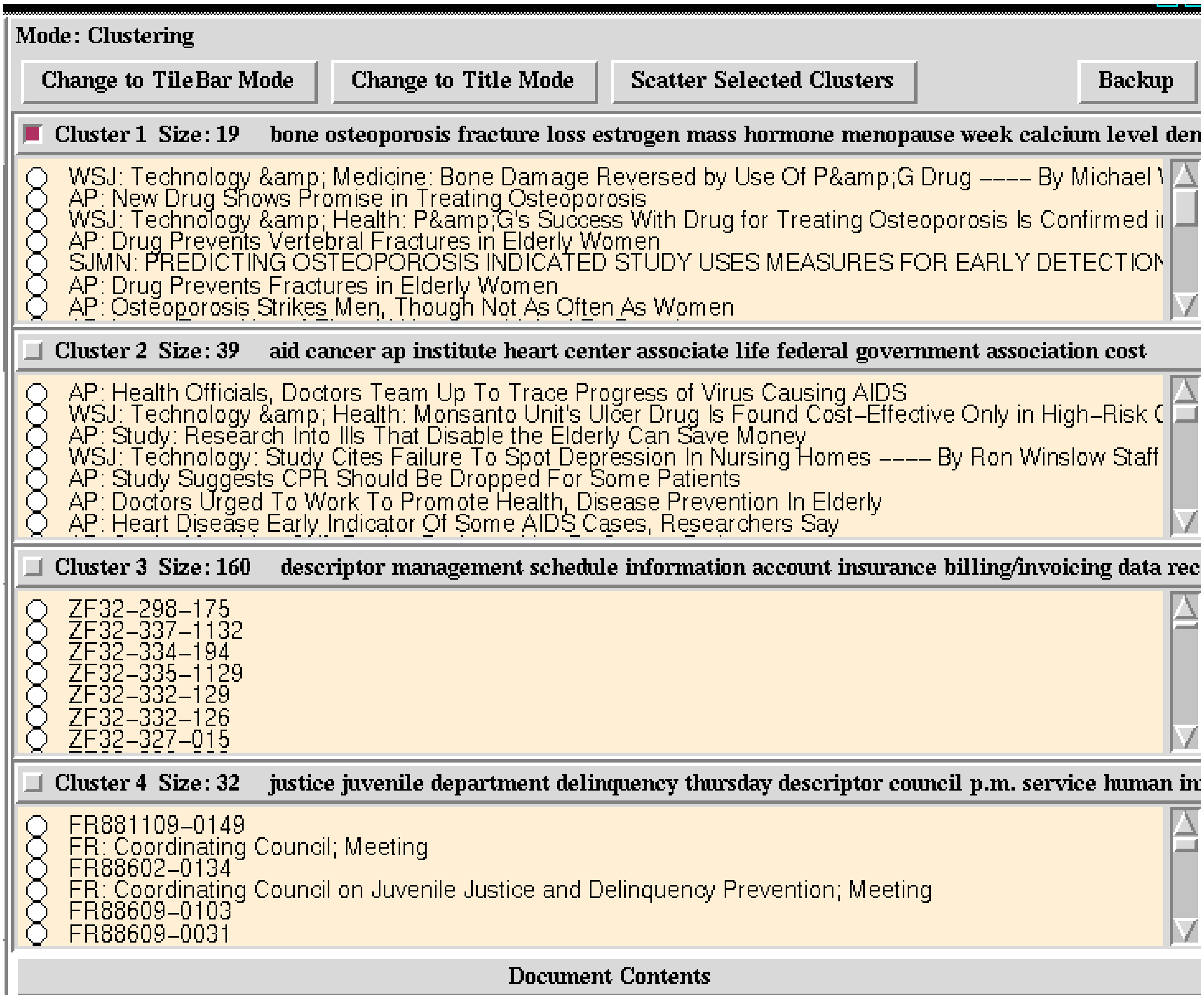

Figure 5: The interface in cluster mode.

A Appendix: Interactive Track System Description

A.1 System Description

The Interactive Track system is described in the body of this report.

A.2 Experimental Conditions

The experimental conditions are described to a large extent in the body of this report. The required information is repeated here.

1. Searcher characteristics
   a. Number of searchers in experiment:
      4
   b. Number of searchers per topic:
      2
   c. Age/age group of searchers:
      20 – 45
   d. IR searching experience of searchers.
      Familiarity with online bibliographic catalogs.
   e. Educational level of searchers.
      Bachelors and Masters degrees
   f. Undergraduate major of searchers.
      Not known.
   g. Experience/familiarity with subject of topic.
      Not known.
   h. Work affiliation of searchers.
      UC Berkeley graduate students.

2. Task description

   Essentially, the description suggested by the leaders of the TREC interactive track: *Find as many good documents as you can for a topic, in around 30 minutes, without collecting too much rubbish.*

3. Training

   Participants completed queries in two sessions. The experiments were run in an otherwise empty room with a video camera recording the session. Participants were given an 10-minute demonstration of the interface followed by a 10-minute warmup exersize. They were also provided with a 3 page description of the interface for reference. A binder of topic descriptions was prepared for each participant, with each topic description appeared on the top of a separate page. Participants were not allowed to look at a new topic before the current one was completed.

III. Search process
Note: the numbers below are only for XERINT2.

1. Elapsed time in seconds per search

   Mean: 1563

   Median: 1627

   SD: 293

   Range: 1104 – 1913

2. Number of documents "viewed" in during the search.

   A document is considered viewed if the user explicitly views the document's contents, as opposed to just seeing the title (and optionally, the associated TileBar).

   Mean: 17.6

   Median: 12

   SD: 14.4

   Range: 4 – 61

3. Number of iterations per search.

   A new iteration takes place when a new search is run for a particular query.

   Mean: 2.1

   Median: 2

   SD: 1.5

   Range: 1 – 6

4. Number of terms used in queries.

   Mean: 5.13

   Median: 8

   SD: 9.3

   Range: 4 – 38

5. Use of system features.

   N/A.

6. Number of user errors made per search.

   N/A.

7. Search narrative for topic 236 (and 216).

   See Appendix B.

B Appendix: Transcripts for Selected Topics

The first number shown on almost every line is the time in seconds. Next appears the mode the user was in, one of TITLES, TILEBARS, or CLUSTERS (for Scatter/Gather). In some cases lists of documents appear in the order in which they were ranked, left to right and top to bottom. Only the top few documents' are shown in each mode (except after the initial search, when all are shown). A 1 indicates a document judged relevant and a 0 a document judged irrelevant. When the participant views or selects a document, it's relevance judgment is also shown.

B.1 Transcript for Topic 216

The first example transcript is that of Topic 216, the running example of this paper.

```
22 TITLES    1 Changing mode to TITLES
80 NEWSEARCH  2
Running new search.
Num docs: 250
Termsets: (OSTEOPOROSIS TREATMENT PREVENTION
           RESEARCH)
0 0 0 0 1 0 0 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0


 181 TITLES    3 Changing mode to TITLES
 206 TITLES    3 Visible Contents of Titles
  ((TITLES) (41)
   (0 0 0 0 1 0 0 1 0 1 0 1 0 0 1 1 1 0 1 0 0
    0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1))

 210 CLUSTERS  4 Changing mode to CLUSTERS
 270 CLUSTERS  4 Visible Contents of Cluster 0
   (DIAGNOSIS FAILURE DESCRIPTOR WASTE ROOF
    MANAGEMENT ANALYSIS DIALYSIS WATER
  CONVENTIONAL DATA U.S.)
   (0 0 0 0 0 0 0 0 0 0 0)

 274 CLUSTERS  4 Visible Contents of Cluster 1
   (OFFICE DEPARTMENT JUSTICE JUVENILE CHILD EDUCATION
    DELINQUENCY WASHINGTON BILL SECRETARY AGENCY GRANT)
   (0 0 0 0 0 0 0 0 0 0 0)

 278 CLUSTERS  4 Visible Contents of Cluster 2
   (INSTITUTE COMMITTEE ADMINISTRATION MENTAL A.M.
    BUILD BILLION ROCKVILLE TIME DEVELOPMENT ROOM
    ALCOHOL)
   (0 0 0 0 0 0 0 0 0 0 0)

 280 CLUSTERS  4 Visible Contents of Cluster 3
   (BONE OSTEOPOROSIS CALCIUM AGE LOSS PREVENT
   FRACTURE UNIVERSITY MENOPAUSE ESTROGEN DR. HORMONE)
   (1 1 1 1 1 1 1 1 1 0 1)

 319 CLUSTERS  4 Selecting Cluster 3
 323 TILEBARS  5 Changing mode to TILEBARS
 332 TILEBARS  5 Visible Contents of Tilebars
   (0 0 0 1 1 1 1 1 1 0 0 1 0 1 1 1 1 1)
```

```
 341 TILEBARS  5 Showing doc FR88513-0157 Tile 7
 383 TILEBARS  5 Showing doc FR88513-0157 Tile 9
 421 TILEBARS  5 Selecting doc FR88513-0157 0
 444 TILEBARS  5 Showing doc AP881110-0214 Tile 0
 468 TILEBARS  5 Showing doc AP881128-0187 Tile 0
 505 TILEBARS  5 Selecting doc AP881128-0187 0
 514 TILEBARS  5 Showing doc WSJ900503-0011 Tile 0
 514 TILEBARS  5 Showing doc WSJ900503-0011 Tile 0
 527 TILEBARS  5 Selecting doc WSJ900503-0011 1
 530 TILEBARS  5 Selecting doc AP900712-0031 1
 539 TILEBARS  5 Showing doc SJMN91-06353137 Tile 0
 540 TILEBARS  5 Showing doc SJMN91-06353137 Tile 0
 575 TILEBARS  5 Selecting doc SJMN91-06295004 1
 580 TILEBARS  5 Showing doc WSJ900712-0096 Tile 0
 580 TILEBARS  5 Showing doc WSJ900712-0096 Tile 0
 592 TILEBARS  5 Selecting doc WSJ900712-0096 1
 597 TILEBARS  5 Showing doc AP900319-0222 Tile 0
 598 TILEBARS  5 Showing doc AP900319-0222 Tile 0
 620 TILEBARS  5 Selecting doc AP900319-0222 0
 638 TILEBARS  5 Showing doc ZF32-150-197 Tile 0
 638 TILEBARS  5 Showing doc ZF32-150-197 Tile 0
 668 TILEBARS  5 Selecting doc ZF32-150-197 0
 670 TILEBARS  5 Showing doc AP900502-0083 Tile 0
 670 TILEBARS  5 Showing doc AP900502-0083 Tile 0
 671 TILEBARS  5 Showing doc AP900502-0083 Tile 0
 677 TILEBARS  5 Selecting doc AP900502-0083 1
 682 TILEBARS  5 Showing doc AP900517-0238 Tile 0
 696 TILEBARS  5 Selecting doc AP900517-0238 0
 699 TILEBARS  5 Selecting doc AP881202-0027 1
 703 TILEBARS  5 Selecting doc AP900927-0033 1
 705 TILEBARS  5 Showing doc WSJ911031-0015 Tile 0
 727 TILEBARS  5 Selecting doc WSJ911031-0015 1
 731 TILEBARS  5 Selecting doc AP900503-0015 1
 735 TILEBARS  5 Selecting doc SJMN91-06275157 1
 738 TILEBARS  5 Showing doc AP881212-0266 Tile 0
 748 TILEBARS  5 Selecting doc AP881212-0266 1
 751 TILEBARS  5 Showing doc SJMN91-06326216 Tile 0
 751 TILEBARS  5 Showing doc SJMN91-06326216 Tile 0
 783 TILEBARS  5 Showing doc AP880603-0121 Tile 0
 783 TILEBARS  5 Showing doc AP880603-0121 Tile 0
 848 TILEBARS  5 Showing doc AP900616-0022 Tile 0
 849 TILEBARS  5 Showing doc AP900616-0022 Tile 0
 854 TILEBARS  5 Showing doc AP900616-0022 Tile 1
 865 TILEBARS  5 Selecting doc AP900616-0022 0
 869 TILEBARS  5 Showing doc SJMN91-06111119 Tile 0
 898 TILEBARS  5 Selecting doc AP900405-0179 1
 904 TILEBARS  5 Showing doc SJMN91-06340009 Tile 0
 905 TILEBARS  5 Showing doc SJMN91-06340009 Tile 0
 929 TILEBARS  5 Selecting doc SJMN91-06340009 1
 934 TILEBARS  5 Showing doc SJMN91-06262031 Tile 0
 935 TILEBARS  5 Showing doc SJMN91-06262031 Tile 0
 946 TILEBARS  5 Selecting doc SJMN91-06262031 1
 948 TILEBARS  5 Unselecting doc SJMN91-06262031 1
 952 TILEBARS  5 Showing doc ZF32-353-056 Tile 0
 963 TILEBARS  5 Selecting doc ZF32-353-056 0
 966 TILEBARS  5 Showing doc SJMN91-06074018 Tile 0
 966 TILEBARS  5 Showing doc SJMN91-06074018 Tile 0
 978 TILEBARS  5 Showing doc AP881212-0255 Tile 0
 979 TILEBARS  5 Showing doc AP881212-0255 Tile 0
1027 TILEBARS  5 Backup up state
1028 CLUSTERS  4 Changing mode to CLUSTERS
1037 CLUSTERS  4 Visible Contents of Cluster 0
   (DIAGNOSIS FAILURE DESCRIPTOR WASTE ROOF MANAGEMENT
    ANALYSIS DIALYSIS WATER CONVENTIONAL DATA U.S.)
   (0 0 0 0 0 0 0 0 0 0 0)

1041 CLUSTERS  4 Visible Contents of Cluster 1
   (OFFICE DEPARTMENT JUSTICE JUVENILE CHILD EDUCATION
    DELINQUENCY WASHINGTON BILL SECRETARY AGENCY GRANT)
   (0 0 0 0 0 0 0 0 0 0 0)

1045 CLUSTERS  4 Visible Contents of Cluster 2
   (INSTITUTE COMMITTEE ADMINISTRATION MENTAL A.M.
   BUILD BILLION ROCKVILLE TIME DEVELOPMENT ROOM ALCOHOL)
   (0 0 0 0 0 0 0 0 0 0 0)

1047 CLUSTERS  4 Visible Contents of Cluster 3
   (BONE OSTEOPOROSIS CALCIUM AGE LOSS PREVENT FRACTURE
    UNIVERSITY MENOPAUSE ESTROGEN DR. HORMONE)
   (1 1 1 1 1 1 1 1 1 0 1)

1058 CLUSTERS  4 Selecting Cluster 2
1061 TILEBARS  6 Changing mode to TILEBARS
```

```
1072 TILEBARS  6 Visible Contents of Tilebars
   (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

1077 TILEBARS  6 Showing doc SJMN91-06059075 Tile 0
1118 TITLES    7 Changing mode to TITLES
1132 TITLES    7 Visible Contents of Titles
  ((TITLES) (41)
   (0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0))

1247 NEWSEARCH  8
Running new search.
Num docs: 500
Termsets: (OSTEOPOROSIS TREATMENT PREVENTION
          RESEARCH)
0 0 0 0 1 0 0 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0


1247 TITLES    9 Changing mode to TITLES
1309 TITLES    9 Visible Contents of Titles
  ((TITLES) (41)
   (0 0 0 0 1 0 0 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0
    0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1))

1316 CLUSTERS 10 Changing mode to CLUSTERS
1382 CLUSTERS 10 Visible Contents of Cluster 0
   (COMMITTEE APPLICATION OFFICE DEPARTMENT PUBLIC
    GRANT BUILD A.M. ROOM MENTAL ADVISORY CHILD)
   (0 0 0 0 0 0 0 0 0 0 0)

1387 CLUSTERS 10 Visible Contents of Cluster 1
   (STUDY WOMAN CANCER RESEARCHER BONE DR.
    UNIVERSITY AGE AMERICAN TEST OSTEOPOROSIS
    MEDICINE)
   (0 1 0 1 0 0 1 0 1 0 0)

1394 CLUSTERS 10 Visible Contents of Cluster 2
   (COMPANY TECHNOLOGY STREET PRODUCT WALL MARKET
    PRESIDENT PAGE GOVERNMENT BIOTECHNOLOGY
   SHARE DIRECTOR)
   (0 0 0 1 0 0 0 0 0 0)

1398 CLUSTERS 10 Visible Contents of Cluster 3
   (DESCRIPTOR DATA MANAGEMENT SCHEDULE USER CASE FAILURE
    GENERIC PROCESS INC. HEAT ORDER)
   (0 0 0 0 0 0 0 0 0 0 0)

1407 CLUSTERS 10 Selecting Cluster 1
1410 CLUSTERS 11 Changing mode to CLUSTERS
1425 CLUSTERS 11 Visible Contents of Cluster 0
   (BONE CALCIUM LOSS FRACTURE MENOPAUSE EXERCISE
    ESTROGEN MASS SUPPLEMENT DENSITY DIET INTAKE)
   (1 1 1 1 1 1 1 1 0 1 1)

1427 CLUSTERS 11 Visible Contents of Cluster 1
   (BREAST HEART FUND BILL NIH MONEY COMMITTEE
   BILLION WASHINGTON DIE ANIMAL CURE)
   (0 0 0 0 0 0 0 0 0 0 0)

1430 CLUSTERS 11 Visible Contents of Cluster 2
   (PATIENT PROGRAM VIRUS INFECTION CASE CLINICAL
    IMMUNE PHYSICIAN HOSPITAL INFORMATION
    NUMB GOVERNMENT)
   (0 0 0 0 0 0 0 0 0 0 0)

1431 CLUSTERS 11 Visible Contents of Cluster 3
   (CHILD GROWTH MUSCLE PATIENT DISORDER PILL
   MEDICATION LAW PARENT DYSTROPHY DIAGNOSIS
   INJECTION)
   (1 0 0 0 0 0 0 0 0 0)

1442 CLUSTERS 11 Selecting Cluster 0
1444 TILEBARS 12 Changing mode to TILEBARS
1455 TILEBARS 12 Visible Contents of Tilebars
   (0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 0)

1467 TILEBARS 12 Showing doc WSJ900402-0142 Tile 0
1473 TILEBARS 12 Showing doc WSJ900402-0142 Tile 3
1490 TILEBARS 12 Showing doc ZF207-107-044 Tile 0
1502 TILEBARS 12 Showing doc SJMN91-06136178 Tile 0
1510 TILEBARS 12 Selecting doc SJMN91-06136178 1
1522 TILEBARS 12 Backup up state
1522 CLUSTERS 11 Changing mode to CLUSTERS
1526 CLUSTERS 11 Visible Contents of Cluster 0
   (BONE CALCIUM LOSS FRACTURE MENOPAUSE EXERCISE
    ESTROGEN MASS SUPPLEMENT DENSITY DIET INTAKE)
   (1 1 1 1 1 1 1 1 0 1 1)

1528 CLUSTERS 11 Visible Contents of Cluster 1
   (BREAST HEART FUND BILL NIH MONEY COMMITTEE BILLION
    WASHINGTON DIE ANIMAL CURE)
   (0 0 0 0 0 0 0 0 0 0 0)

1534 CLUSTERS 11 Visible Contents of Cluster 2
   (PATIENT PROGRAM VIRUS INFECTION CASE CLINICAL
    IMMUNE  PHYSICIAN HOSPITAL INFORMATION
    NUMB GOVERNMENT)
   (0 0 0 0 0 0 0 0 0 0 0)

1535 CLUSTERS 11 Visible Contents of Cluster 3
   (CHILD GROWTH MUSCLE PATIENT DISORDER PILL
   MEDICATION LAW PARENT DYSTROPHY DIAGNOSIS
   INJECTION)
   (1 0 0 0 0 0 0 0 0 0)

1536 CLUSTERS 11 Selecting Cluster 1
1540 TILEBARS 13 Changing mode to TILEBARS
1549 TILEBARS 13 Visible Contents of Tilebars
   (0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1)

1559 TILEBARS 13 Showing doc AP901106-0019 Tile 2
1572 TILEBARS 13 Selecting doc AP901106-0019 1
1576 TILEBARS 13 Showing doc AP900502-0035 Tile 1
1603 TILEBARS 13 Showing doc WSJ910403-0155 Tile 5
1610 TILEBARS 13 Showing doc WSJ910403-0155 Tile 5
1627 TILEBARS 13 Selecting doc WSJ910403-0155 1
1631 TILEBARS 13 Showing doc SJMN91-06175092 Tile 0
1651 TILEBARS 13 Selecting doc SJMN91-06175092 1
1664 TILEBARS 13 Showing doc SJMN91-06326216 Tile 2
1684 TILEBARS 13 Showing doc WSJ910619-0089 Tile 1
1979 TILEBARS 13 Query done

Final selected documents:

(FR88513-0157 0) (AP881128-0187 0)
(WSJ900503-0011 1) (AP900712-0031 1)
(SJMN91-06295004 1) (WSJ900712-0096 1)
 (AP900319-0222 0) (ZF32-150-197 0)
(AP900502-0083 1) (AP900517-0238 0)
(AP881202-0027 1) (AP900927-0033 1)
(WSJ911031-0015 1) (AP900503-0015 1)
(SJMN91-06275157 1) (AP881212-0266 1)
(AP900616-0022 0) (AP900405-0179 1)
(SJMN91-06340009 1) (ZF32-353-056 0)
(SJMN91-06136178 1) (AP901106-0019 1)
(WSJ910403-0155 1) (SJMN91-06175092 1)
```

B.2 Transcript for Topic 236

As with most of the Interactive Track systems at TREC4, the results on Topic 236 were not the strongest. In this case, only one search was run. Of the 8 possible relevant documents from the search, 6 of these were found and saved.

```
23 TITLES    1 Changing mode to TITLES
311 NEWSEARCH  2
Running new search.
Num docs: 500
Termsets: ({LAW LEGAL REGULATION} {SEA OCEAN} COMPARE
          {AGREEMENT DISAGREEMENT})
0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0


 314 TITLES    3 Changing mode to TITLES
 452 TITLES    3 Visible Contents of Titles
  ((TITLES) (41)
   (0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0))

 463 CLUSTERS  4 Changing mode to CLUSTERS
 578 CLUSTERS  4 Visible Contents of Cluster 0
   (WATER AP DUMP SOVIET MARINE MILE COAST OIL NATIONAL
    WORLD VESSEL SITE)
  (0 1 0 0 0 0 0 0 1 0 0)

 587 CLUSTERS  4 Visible Contents of Cluster 1
   (FUND CC AVAILABILITY SCHEDULE CHECK APPLY DEPOSIT
    CHILD ACCOUNT AUTHORITY INSTITUTION LEA)
   (0 0 0 0 0 0 0 0 0 0 0)

 753 CLUSTERS  4 Visible Contents of Cluster 2
   (LEGAL COURT FIRM MR. LAWYER CASE JUSTICE JUDGE WALL
    SECURITY GOVERNMENT ATTORNEY)
   (0 0 0 0 0 0 0 0 0 0 0)

 773 CLUSTERS  4 Visible Contents of Cluster 3
   (COMMISSION PARTY MARITIME TITLE LINE INTEREST
   SYNOPSIS PORT DC FREIGHT LICENSE CONFERENCE)
   (0 0 0 0 0 0 0 0 0 0 0)

 813 CLUSTERS  4 Selecting Cluster 3
 817 CLUSTERS  4 Selecting Cluster 0
 823 CLUSTERS  5 Changing mode to CLUSTERS
 846 CLUSTERS  5 Visible Contents of Cluster 0
   (FOOT RESCUE GUARD SINK ABOARD CREW GAS SAFETY AIR
    CATTLE SPILL SOUTH)(0 0 0 0 0 0 0 0 0 0 0)

 856 CLUSTERS  5 Visible Contents of Cluster 1
   (SOVIET FISH WORLD TIME FISHERY SCIENTIST COUNTRY
    BERING UNION LION PACIFIC NATION)
   (1 1 0 0 0 0 0 1 0 0 0)

 877 CLUSTERS  5 Visible Contents of Cluster 2
   (COMMISSION PARTY MARITIME FILE TITLE LINE
    SECTION INTEREST SYNOPSIS CODE PORT SECRETARY)
   (0 0 0 0 0 0 0 0 0 0 0)

 887 CLUSTERS  5 Visible Contents of Cluster 3
   (DUMP SITE WASTE DISPOSAL DRILL SLUDGE EPA
    OFFER YORK ENVIRONMENTAL BAN COMPANY)
   (0 0 0 0 0 0 0 0 0 0 0)

 889 CLUSTERS  5 Showing doc FR88512-0056 Tile 0
 956 CLUSTERS  5 Showing doc FR881018-0063 Tile 0
 989 CLUSTERS  5 Showing doc FR88607-0070 Tile 0
1017 CLUSTERS  5 Selecting Cluster 0
1020 CLUSTERS  5 Selecting Cluster 3
1025 CLUSTERS  5 Selecting Cluster 1
1030 CLUSTERS  6 Changing mode to CLUSTERS
1051 CLUSTERS  6 Visible Contents of Cluster 0
   (CONTAINER GAS TEMPLE DEAD FERRY PETROLEUM
   SCROLL BID RIG GOVERNMENT ASSET EXPLOSION)
   (0 0 0 0 0 0 0 0 0 0 0)

1065 CLUSTERS  6 Visible Contents of Cluster 1
   (U.S. FISH SOVIET SHIP UNITE OFFICIAL VESSEL
    WORLD TIME WASHINGTON SCIENTIST ISLAND)
   (1 0 0 0 1 0 0 1 0 0 0)

1067 CLUSTERS  6 Visible Contents of Cluster 2
   (WASTE SLUDGE BAN INDUSTRIAL BILL POLLUTION
   JERSEY MEDICAL SEWAGE HOUSE SENATE CONGRESS)
   (0 0 0 0 0 0 0 0 0 0 0)

1071 CLUSTERS  6 Visible Contents of Cluster 3
   (DRILL MURPHY MANAGEMENT DREDGE MATERIAL IMPACT
    LA-5 DEEP ADMINISTRATION ACTION
    DESIGNATION PROGRAM)
   (0 0 0 0 0 0 0 0 0 0 0)

1074 CLUSTERS  6 Selecting doc AP880810-0056 0
1076 CLUSTERS  6 Selecting doc AP880809-0150 0
1080 CLUSTERS  6 Selecting doc AP881118-0130 0
1084 CLUSTERS  6 Selecting doc AP881018-0197 0
1086 CLUSTERS  6 Selecting doc AP881008-0018 0
1089 CLUSTERS  6 Selecting doc AP901029-0103 0
1101 CLUSTERS  6 Selecting doc AP901103-0030 0
1102 CLUSTERS  6 Unselecting doc AP901103-0030 0
1102 CLUSTERS  6 Selecting doc AP901103-0030 0
1113 CLUSTERS  6 Selecting doc FR881129-0025 0
1116 CLUSTERS  6 Selecting doc SJMN91-06011130 0
1120 CLUSTERS  6 Selecting doc AP881213-0090 0
1133 CLUSTERS  6 Selecting doc AP880830-0087 0
1140 CLUSTERS  6 Selecting doc SJMN91-06254147 0
1141 CLUSTERS  6 Selecting doc AP900404-0116 0
1158 CLUSTERS  6 Selecting doc WSJ910618-0147 0
1159 CLUSTERS  6 Selecting doc WSJ910124-0116 0
1197 CLUSTERS  6 Selecting Cluster 3
1201 CLUSTERS  6 Selecting Cluster 1
1209 CLUSTERS  6 Selecting Cluster 0
1213 CLUSTERS  7 Changing mode to CLUSTERS
1226 CLUSTERS  7 Visible Contents of Cluster 0
   (OIL DRILL EXPLORATION GAS MURPHY PETROLEUM
   ENERGY JOURNAL WALL STREET PET PAGE)
   (0 0 0 0 0 0 0 0 0 0 0)

1233 CLUSTERS  7 Visible Contents of Cluster 1
   (SITE DISPOSAL MANAGEMENT DEEP IMPACT DREDGE
    MATERIAL LAVA PROGRAM LA-5 HOME ADMINISTRATION)
   (0 0 0 0 0 0 0 0 0 0 0)

1249 CLUSTERS  7 Visible Contents of Cluster 2
   (U.S. SOVIET SHIP UNITE VESSEL OFFICIAL WASHINGTON
    ISLAND FISHERY INTERNATIONAL COUNTRY BERING)
   (1 1 0 0 0 0 0 1 0 0 0)

1256 CLUSTERS  7 Visible Contents of Cluster 3
   (CONTAINER TEMPLE SCROLL FERRY SALE DEAD ASSET
    STENA BILLION BUY TIPHOOK RECAPITALIZATION)
   (0 0 0 0 0 0 0 0 0 0 0)

1257 CLUSTERS  7 Selecting Cluster 0
1259 TILEBARS  8 Changing mode to TILEBARS
1271 TILEBARS  8 Visible Contents of Tilebars
   (0 0 0 0 0 0 0 0 0 0 0 0 0)

1278 TILEBARS  8 Selecting doc SJMN91-06108037 0
1284 TILEBARS  8 Unselecting doc WSJ910124-0116 0
1289 TILEBARS  8 Unselecting doc WSJ910618-0147 0
1290 TILEBARS  8 Selecting doc WSJ910618-0147 0
```

```
1290 TILEBARS  8 Unselecting doc WSJ910618-0147 0
1299 TILEBARS  8 Selecting doc AP880930-0025 0
1301 TILEBARS  8 Selecting doc AP880929-0298 0
1305 TILEBARS  8 Backup up state
1305 CLUSTERS  7 Changing mode to CLUSTERS
1310 CLUSTERS  7 Visible Contents of Cluster 0
   (OIL DRILL EXPLORATION GAS MURPHY PETROLEUM
   ENERGY JOURNAL WALL STREET PET PAGE)
   (0 0 0 0 0 0 0 0 0 0 0)

1314 CLUSTERS  7 Visible Contents of Cluster 1
   (SITE DISPOSAL MANAGEMENT DEEP IMPACT DREDGE
    MATERIAL LAVA PROGRAM LA-5 HOME ADMINISTRATION)
   (0 0 0 0 0 0 0 0 0 0 0)

1324 CLUSTERS  7 Visible Contents of Cluster 2
   (U.S. SOVIET SHIP UNITE VESSEL OFFICIAL WASHINGTON
   ISLAND FISHERY INTERNATIONAL COUNTRY BERING)
   (1 1 0 0 0 0 0 1 0 0 0)

1326 CLUSTERS  7 Visible Contents of Cluster 3
   (CONTAINER TEMPLE SCROLL FERRY SALE DEAD ASSET
   STENA BILLION BUY TIPHOOK RECAPITALIZATION)
   (0 0 0 0 0 0 0 0 0 0 0)

1327 CLUSTERS  7 Selecting Cluster 3
1330 TILEBARS  9 Changing mode to TILEBARS
1347 TILEBARS  9 Visible Contents of Tilebars
   (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

1358 TILEBARS  9 Backup up state
1359 CLUSTERS  7 Changing mode to CLUSTERS
1362 CLUSTERS  7 Visible Contents of Cluster 0
   (OIL DRILL EXPLORATION GAS MURPHY PETROLEUM
    ENERGY JOURNAL WALL STREET PET PAGE)
   (0 0 0 0 0 0 0 0 0 0 0)

1364 CLUSTERS  7 Visible Contents of Cluster 1
   (SITE DISPOSAL MANAGEMENT DEEP IMPACT DREDGE
   MATERIAL LAVA PROGRAM LA-5 HOME ADMINISTRATION)
   (0 0 0 0 0 0 0 0 0 0 0)

1371 CLUSTERS  7 Visible Contents of Cluster 2
   (U.S. SOVIET SHIP UNITE VESSEL OFFICIAL WASHINGTON
   ISLAND FISHERY INTERNATIONAL COUNTRY BERING)
   (1 1 0 0 0 0 0 1 0 0 0)

1372 CLUSTERS  7 Visible Contents of Cluster 3
   (CONTAINER TEMPLE SCROLL FERRY SALE DEAD ASSET
    STENA BILLION BUY TIPHOOK RECAPITALIZATION)
   (0 0 0 0 0 0 0 0 0 0 0)

1372 CLUSTERS  7 Selecting Cluster 1
1374 CLUSTERS  7 Selecting Cluster 2
1378 CLUSTERS 10 Changing mode to CLUSTERS
1392 CLUSTERS 10 Visible Contents of Cluster 0
   (SITE FOOT DISPOSAL CATTLE DEEP IMPACT PEN DREDGE
    MATERIAL SQUARE RULE LA-5)
   (0 0 0 0 0 0 0 0 0 0 0)

1394 CLUSTERS 10 Visible Contents of Cluster 1
   (SOVIET AGREEMENT BERING COUNTRY UNION INTERNATIONAL
   AMERICAN NATION SIGN FOREIGN MOSCOW ZONE)
   (1 1 0 0 0 0 1 0 0 0 0)

1401 CLUSTERS 10 Visible Contents of Cluster 2
   (LION MAMMAL OTTER CALIFORNIA BOAT ANIMAL
    HOUSE CITY SEAL FEDERAL KILL COUNTY)
   (0 0 0 0 0 0 0 0 0 0 0)

1406 CLUSTERS 10 Visible Contents of Cluster 3
   (RESCUE GUARD OIL CREW ABOARD SINK SOUTH
    HELICOPTER AIRCRAFT SEARCH FUEL CRUISE)
   (0 0 0 0 0 0 0 0 0 0 0)

1423 CLUSTERS 10 Showing doc AP900527-0035 Tile 0
1429 CLUSTERS 10 Selecting Cluster 0
1431 TILEBARS 11 Changing mode to TILEBARS
1438 TILEBARS 11 Visible Contents of Tilebars
   (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

1443 TILEBARS 11 Selecting doc FR88826-0028 0
1448 TILEBARS 11 Selecting doc FR881017-0004 0
1454 TILEBARS 11 Selecting doc SJMN91-06151045 0
1463 TILEBARS 11 Selecting doc SJMN91-06263151 0
1481 TILEBARS 11 Backup up state
1481 CLUSTERS 10 Changing mode to CLUSTERS
1484 CLUSTERS 10 Visible Contents of Cluster 0
   (SITE FOOT DISPOSAL CATTLE DEEP IMPACT PEN DREDGE
   MATERIAL SQUARE RULE LA-5)
  (0 0 0 0 0 0 0 0 0 0 0)

1492 CLUSTERS 10 Visible Contents of Cluster 1
   (SOVIET AGREEMENT BERING COUNTRY UNION INTERNATIONAL
    AMERICAN NATION SIGN FOREIGN MOSCOW ZONE)
   (1 1 0 0 0 0 1 0 0 0 0)

1493 CLUSTERS 10 Visible Contents of Cluster 2
   (LION MAMMAL OTTER CALIFORNIA BOAT ANIMAL
    HOUSE CITY SEAL FEDERAL KILL COUNTY)
   (0 0 0 0 0 0 0 0 0 0 0)

1494 CLUSTERS 10 Visible Contents of Cluster 3
   (RESCUE GUARD OIL CREW ABOARD SINK SOUTH
    HELICOPTER AIRCRAFT SEARCH FUEL CRUISE)
   (0 0 0 0 0 0 0 0 0 0 0)

1495 CLUSTERS 10 Selecting Cluster 1
1498 TILEBARS 12 Changing mode to TILEBARS
1511 TILEBARS 12 Visible Contents of Tilebars
   (0 1 0 0 1 0 0 1 1 0 0 0 1 0 0 0 0 0)

1513 TILEBARS 12 Selecting doc FR88616-0064 0
1517 TILEBARS 12 Selecting doc AP881101-0176 1
1523 TILEBARS 12 Selecting doc AP900601-0165 0
1528 TILEBARS 12 Selecting doc AP900212-0157 0
1538 TILEBARS 12 Selecting doc AP881111-0203 0
1542 TILEBARS 12 Selecting doc AP880419-0033 1
1545 TILEBARS 12 Selecting doc AP900831-0147 0
1552 TILEBARS 12 Selecting doc AP881025-0139 1
1556 TILEBARS 12 Selecting doc AP880518-0301 1
1563 TILEBARS 12 Selecting doc AP880817-0026 0
1579 TILEBARS 12 Selecting doc AP880423-0166 0
1582 TILEBARS 12 Selecting doc AP881024-0151 0
1586 TILEBARS 12 Selecting doc AP880425-0174 0
1589 TILEBARS 12 Selecting doc AP900129-0102 0
1596 TILEBARS 12 Unselecting doc AP900129-0102 0
1599 TILEBARS 12 Selecting doc AP881228-0105 1
1605 TILEBARS 12 Selecting doc AP900402-0131 0
1612 TILEBARS 12 Selecting doc AP900109-0043 1
1640 TILEBARS 12 Backup up state
1640 CLUSTERS 10 Changing mode to CLUSTERS
1647 CLUSTERS 10 Visible Contents of Cluster 0
   (SITE FOOT DISPOSAL CATTLE DEEP IMPACT PEN
    DREDGE MATERIAL SQUARE RULE LA-5)
   (0 0 0 0 0 0 0 0 0 0 0)

1653 CLUSTERS 10 Visible Contents of Cluster 1
   (SOVIET AGREEMENT BERING COUNTRY UNION
    INTERNATIONAL AMERICAN NATION SIGN FOREIGN
    MOSCOW ZONE)
   (1 1 0 0 0 0 1 0 0 0 0)

1655 CLUSTERS 10 Visible Contents of Cluster 2
   (LION MAMMAL OTTER CALIFORNIA BOAT ANIMAL
    HOUSE CITY SEAL FEDERAL KILL COUNTY)
   (0 0 0 0 0 0 0 0 0 0 0)

1656 CLUSTERS 10 Visible Contents of Cluster 3
   (RESCUE GUARD OIL CREW ABOARD SINK SOUTH
    HELICOPTER AIRCRAFT SEARCH FUEL CRUISE)
   (0 0 0 0 0 0 0 0 0 0 0)

1658 CLUSTERS 10 Selecting Cluster 2
1665 CLUSTERS 10 Selecting Cluster 3
1668 TILEBARS 13 Changing mode to TILEBARS
1684 TILEBARS 13 Visible Contents of Tilebars
   (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

1686 TILEBARS 13 Selecting doc FR88927-0030 0
1701 TILEBARS 13 Selecting doc AP900517-0140 0
1730 TILEBARS 13 Selecting doc SJMN91-06212079 0
1732 TILEBARS 13 Selecting doc SJMN91-06219052 0
1742 TILEBARS 13 Selecting doc AP900419-0113 0
1867 TILEBARS 13 Query done
```

```
Final selected documents:

(AP880810-0056 0) (AP880809-0150 0)
(AP881118-0130 0) (AP881018-0197 0)
(AP881008-0018 0) (AP901029-0103 0)
(AP901103-0030 0) (FR881129-0025 0)
(SJMN91-06011130 0) (AP881213-0090 0)
(AP880830-0087 0) (SJMN91-06254147 0)
(AP900404-0116 0) (SJMN91-06108037 0)
(AP880930-0025 0) (AP880929-0298 0)
(FR88826-0028 0) (FR881017-0004 0)
(SJMN91-06151045 0) (SJMN91-06263151 0)
(FR88616-0064 0) (AP881101-0176 1)
(AP900601-0165 0) (AP900212-0157 0)
(AP881111-0203 0) (AP880419-0033 1)
(AP900831-0147 0) (AP881025-0139 1)
(AP880518-0301 1) (AP880817-0026 0)
(AP880423-0166 0) (AP881024-0151 0)
(AP880425-0174 0) (AP881228-0105 1)
(AP900402-0131 0) (AP900109-0043 1)
(FR88927-0030 0) (AP900517-0140 0)
(SJMN91-06212079 0) (SJMN91-06219052 0)
(AP900419-0113 0)
```