Information Retrieval 1, 7–34 (1999)

Blind Men and Elephants: Six Approaches to TREC data

DAVID BANKS, PAUL OVER, AND NIEN-FAN ZHANG
National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899

Received July 8, 1998; Revised October 8, 1998; Accepted October 9, 1998

Abstract. The paper reviews six recent efforts to better understand performance measurements on information retrieval (IR) systems within the framework of the Text REtrieval Conferences (TREC): analysis of variance, cluster analyses, rank correlations, beadplots, multidimensional scaling, and item response analysis. None of this work has yielded any substantial new insights. Prospects that additional work along these lines will yield more interesting results vary but are in general not promising. Some suggestions are made for paying greater attention to richer descriptions of IR system behavior but within smaller, better controlled settings.

Keywords: cluster analysis, multidimensional scaling, rank correlation

1. The TREC problem

Usually one writes papers about problems that are solved. But sometimes problems are so difficult that an examination of the analyses that failed, and why they failed, is valuable. Thomas Edison, while working on the storage battery, was asked by his friend W.S. Mallory "Isn't it a shame that with the tremendous amount of work that you've done, you've not been able to get any results?" To which Edison replied, "Results! Why man, I have gotten a lot of results. I know several thousand things that won't work" (Dyer and Martin 1910).

The analysis of the interplay of factors affecting the performance of information retrieval systems is a hard problem (cf. Lawrence and Giles 1998). In the spirit of Edison's response, we hope this review of six recent efforts to better understand performance measurements on information retrieval systems within the framework of the Text REtrieval Conference (TREC) collections may help other researchers.

TREC workshops began in 1992, and provide information retrieval researchers with (1) large test collections of documents, (2) benchmark structured statements of information need (topics) to study, and (3) a forum in which to compare results. To date, TREC has issued five compact disks of English documents (about 5 gigabytes of text) and 350 retrieval topics. Background on TREC can be found at `http://trec.nist.gov`, which includes publications providing an overview of each TREC conference's framework, processes, and results as well as information on the topics and documents in each year's test collection and detailed information about the work of participating groups.

The process of creating TREC test collections has been relatively unchanged since TREC-3 in 1994:

- Relevance assessors (retired information analysts) each propose ten candidate topics, and search the target collection for related documents that meet their personal criteria for topic relevance. Based on the estimated number of relevant documents and load balancing across the assessors, fifty topics are chosen.
- The chosen topics are used in that year's TREC. Participating researchers use their information retrieval systems to search the target collection for each topic, and then submit a ranked list of the 1000 most probably relevant documents to NIST for evaluation. Such a list is called a run. A participant may submit runs from two or three variants of their system.
- For each topic, NIST pools the top 100 retrieved documents from each run. The relevance assessor who proposed the topic then examines each document and makes a binary determination of 'relevant' or 'not relevant'. All documents not in the pool are assumed to be irrelevant.

It is notable that when the pool is assessed by judges other than the original topic proposer, fewer documents are deemed relevant. But the topic proposer is the model user, and thus the ultimate arbiter. Nonetheless, this subjective criterion makes objective evaluation hard.

The topics are chosen according to several criteria, but a consistent aim has been to make the retrieval task similar to a typical library search. To illustrate this, the topics that are separately analyzed in this paper appear in Table 1. These topics span a range of difficulty. Topic 341, on Airport Security, is fairly easy—two proximate keywords give good performance. Topic 326, on Ferry Sinkings, is slightly above average in difficulty because it is difficult to codify the requirement that the death toll exceeds 100. Topic 336, on Black Bear Attacks, is hard, probably because all three keywords are required for specificity.

In TREC, each retrieval system is given the list of topics and expanded narrative explanations, then asked to search a large document collection, ranking the documents according to apparent relevance. For example, in TREC-6 (1997) the document collection comprised about one gigabyte of news articles, newswire data, computer-related articles, and government documents from the following five databases:

- *Financial Times of London* articles from 1991 to 1994.
- *Federal Register* documents from 1994.
- Documents from *Congressional Record* (*103rd*).
- Articles from *FBI Service*, 1993.
- Articles from the L.A. Times, 1993–1994.

Note that some of these articles will be extremely similar, perhaps even identical.

The evaluation of the systems by NIST is based upon several different measures of recall and/or precision. Recall assesses the fraction of relevant documents in the collection that were found by a system, while precision assesses the fraction of a system's retrieved documents that are actually relevant—these ideas are closely related to Type I and Type II error in statistics, or to sensitivity and specificity in medicine. Essentially, each system must balance the risk of excluding a relevant document against the risk of including an irrelevant one.

Table 1. TREC-6 topics that are discussed individually.

Topic 322: International Art Crime.

Description: Isolate instances of fraud or embezzlement in the international art trade.
Narrative: A relevant document is any report that identifies an instance of fraud or embezzlement in the international buying or selling of art objects. Objects include paintings, jewelry, sculptures and any other valuable works of art. Specific instances must be identified for a document to be relevant; generalities are not relevant.

Topic 326: Ferry Sinkings.

Description: Any report of a ferry sinking where 100 or more people lost their lives.
Narrative: To be relevant, a document must identify a ferry that has sunk causing the death of 100 or more humans. It must identify the ferry by name or place where the sinking occurred. Details of the cause of the sinking would be helpful but are not necessary to be relevant. A reference to a ferry sinking without the number of deaths would not be relevant.

Topic 336: Black Bear Attacks.

Description: A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.
Narrative: It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

Topic 341: Airport Security.

Description: A relevant document would discuss how effective government orders to better scrutinize passengers and luggage on international flights and to step up screening of all carry-on baggage has been.
Narrative: A relevant document would contain reports on what new steps airports worldwide have taken to better scrutinize passengers and their luggage on international flights and to step up screening of all carry-on baggage. With the increase in international terrorism and in the wake of the TWA Flight 800 disaster, articles on airport security relating in particular to additional steps taken by airports to increase flight safety would be relevant. The mere mention of enhanced security does not constitute relevance. Additional steps refer to something beyond just passenger and carry-on screening using the normal methods. Examples of new steps would be additional personnel, sophisticated monitoring and screening devices, and extraordinary measures to check luggage in the baggage compartment.

The primary purpose of this paper is to describe analyses that can help discover whether there are real differences among the retrieval systems, and to understand how such differences relate to differences among topics and collections of documents. The specific performance measures used in this paper are *average precision* and *overlap*, which are described in the following sections. With these measures, the following six sections examine six strategies for analyzing TREC data:

Analysis of Variance, which looks for sources of variability in retrieval performance that can be attributed to system, topic, or interaction between the retrieval system and the topic in terms of performance;

Cluster Analyses, which look for cluster structure among topics, systems, and both together;

Rank Correlations, which exploit the ordering of the documents to measure similarity among the systems;

Beadplots, a new visualization tool that emphasizes shared subpatterns in the sequence of retrieved documents;

Multidimensional Scaling, to determine whether system performance is well-described in some low-dimensional space; and

Item Response Analysis, which uses latent variable theory to exploit an analogy between students taking tests and systems retrieving documents.

Several of these approaches contain multiple parts; e.g., we examine different kinds of analysis of variance and clustering. Data from TREC-6 is used throughout except for the analysis of variance, which extends previous work done on TREC-3 data.

Ultimately, the purpose of our work is to reconstruct the elephant. We want to understand the document retrieval problem well enough to know which kinds of retrieval strategies work best for which topics, and which performance differences among systems are real, rather than noise. This quest is complicated by the fact that there is no single measure of performance that applies to all situations, and by the fact that the retrieval strategies tend to be complex accretions of tactics.

2. Analysis of variance

A two-way analysis of variance without interaction for TREC-3 data has been done by Tague-Sutcliffe and Blustein (1995). They used the average precision Y_{ij} of the ith system on the jth topic as the response, with average precision is defined as

$$Y_{ij} = R^{-1} \sum_{d \in D} \frac{\#\{\text{relevant documents retrieved at or before document } d\}}{\#\{\text{number of documents retrieved at or before document } d\}} I(d)$$

where R is the number of relevant documents, D is the set of all documents, $\#\{\cdot\}$ denotes the cardinality of the argument set, and $I(d)$ is an indicator function which takes the value 1 if d is a relevant document, and otherwise is 0. This measure was proposed by Harman (1994), and it has been widely employed in TREC literature (cf. Harman 1996). One can view this as the area under the step-function representing the cumulative proportion of relevant documents found in the ranked list from system i on topic j.

To the average precision of the 42 retrieval systems used in TREC-3 over 50 topics, Tague-Sutcliffe and Blustein fit the ANOVA model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, 42, \;\; j = 1, 50.$$

Here α_i is the effect due to the system, β_j is the effect due to the topic, and ϵ_{ij} captures all other variation. This model assumes that the factors of system and topic are additive, which is equivalent to asserting that there is no interactive effect between system and topic. Table 2 shows the ANOVA table obtained in her analysis. The proportion of variation explained by the factor effects (or R^2, the coefficient of determination) is .738.

Table 2. Tague-Sutcliffe and Blustein's analysis of variance table, using an additive model fitted to TREC-3 data. Here $R^2 = 0.738$.

Source	Degrees of freedom	Sum of squares	Mean square	F	P-value
System	41	15.42	0.376	34.44	0.0001
Topic	49	46.25	0.944	86.46	0.0001
Error	2009	21.93	0.011		
Total	2099	83.60			

From Table 2, it is clear that both system and topic affect the average precision. But this analysis does not speak to interaction, which occurs when the effect of one level of a factor depends on the level of another factor. In this case, interaction would be present if some systems were better for some topics, but were systematically worse for others.

To assess interaction, we need to fit a more complex model. This is difficult because we have only one observation for each combination of the levels of topic and system, which precludes a general interaction analysis. Following Milliken and Johnson (1991), we examine two models that are designed to capture specific forms of interaction, rather than all possible interactions. One is a test that allocates a single degree of freedom to estimating interactions, while the other is more flexible, fitting an effect for each system. Neither model can capture all kinds of interaction; however, if they find interaction, then it is surely present, though probably in a more complicated form than the fitted model indicates.

The first model we consider uses Tukey's single-degree-of-freedom test for nonadditivity. This assumes the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j + \epsilon_{ij}$$

where the interaction term takes a specific multiplicative form. The corresponding ANOVA table is shown in Table 3. The results show strongly significant interaction, in the third row of the table. Note that the new value of $R^2 = 0.768$ indicates slightly better fit.

Table 3. Tukey's Single-Degree-of-Freedom test for interaction, using an analysis of variance model fitted to TREC-3 data. Here $R^2 = 0.768$.

Source	Degrees of freedom	Sum of squares	Mean square	F	P-value
System	41	15.42	0.376	39.00	0.0001
Topic	49	46.25	0.944	97.91	0.0001
Interaction	1	2.57	2.570	266.93	0.0001
Error	2008	19.36	0.010		
Total	2099	83.60			

Table 4. Mandel's Bundle-of-Lines test for interaction, using an analysis of variance model fitted to TREC-3 data. Here $R^2 = 0.782$.

Source	Degrees of freedom	Sum of squares	Mean square	F	P-value
System	41	15.42	0.376	40.41	0.0001
Topic	49	46.25	0.944	101.45	0.0001
Interaction	50	3.70	0.074	7.96	0.0001
Error	1959	18.22	0.009		
Total	2099	83.60			

The second model uses Mandel's bundle-of-lines approach, which generalizes Tukey's model by fitting

$$Y_{ij} = \mu + \alpha_i + \beta_j + \lambda\delta_i\beta_j + \epsilon_{ij}.$$

Geometrically, this corresponds to capturing the interaction as nonparallel slopes in the topic effect. Table 4 shows the resulting ANOVA table. Again, the interaction term is highly significant, and the fit improves slightly upon Tukey's model, as reflected in the R^2 value.

When the performance criterion is average precision, there is significant interaction between the systems and the topics. The magnitude of this interaction is unclear, since the structure of the problem (no replicates for system-topic pairs) makes full modelling of the interaction term impossible. The best we can do is to fit specific forms of interaction, which provides lower bounds on the effect. We note that Mandel's more flexible test offers little improvement in fit over that obtained from Tukey's test, as measured by R^2, and this suggests that although interaction is clearly present, it may be well-approximated by some simple form—this bears further investigation. Similarly, Tukey's R^2 is not much larger than that obtained by Tague-Sutcliffe and Blustein hinting that although statistically significant, interactive effects may be small compared to the additive effects.

Sometimes it happens that all of the interaction in a dataset is due to a few combinations of the factor levels, and that most combinations have additive structure. If those non-additive cases can be identified and removed, then a simple model can be fit to the remaining ones. In the case of the TREC-3 data, there is a set of 13 topics among the 50 (numbers 151, 154, 156, 161, 163, 166, 170, 173, 174, 185, 193, 196, and 198) for which no significant interaction with the systems is found by either Tukey's or Mandel's tests. This result suggests these topics may have a usefully simple relationship with the systems in terms of average precision, although it does not preclude the possibility that all 13 cases have interactions not measurable by the models of Tukey or Mandel.

The results from these studies show there is a strong interaction between system and topic in terms of average precision. The presence of interaction implies that one cannot find simple descriptions of the data in terms of topics and systems alone. Early hopes that simpler structure would obtain, permitting definitive rankings of systems or topics, were too optimistic. It is conceivable that some other measure of system performance than

average precision might enjoy additive structure, but such measures would be coarser and less informative.

3. Cluster analysis

Domain experts believe some kind of cluster structure exists in the TREC data, among the topics, the retrieval systems, and the documents. Given the interactions found in the previous analysis, it seems best to look first for cluster structure within topics, thereby avoiding the complication caused by different system behavior across topics. In that spirit, we first performed block-cluster analyses of systems and documents within topics.

Heuristically, the block-cluster analysis constructs a matrix whose rows represent documents and whose columns represent systems. The entries in the matrix are the rank in which the document was retrieved by the system. Then the analysis permutes the rows and columns so as to minimize the differences among the neighboring entries in the matrix. This produces 'blocks' of similar values within the matrix, and also can often induce interpretable cluster structure among the rows (documents) and columns (systems).

In reality, the block-clustering algorithm and our instantiation of it are slightly more complex. Regarding our instantiation, rather than use the actual rank of each document, we have replaced ranks larger than 99 with the value 99 (this ensures the algorithm doesn't churn pointlessly, trying to make block distinctions among degrees of irrelevance that have little practical importance). Also, the complexity of the computational problem increases dramatically with the numbers of rows and columns. Thus we restrict attention to 23 systems (chosen by TREC experts to include generally good systems that span a range of search strategies and use both manually-tuned and purely automatic techniques) and 50 documents. The 50 documents represent the first 50 chosen by any of the systems; thus we first include all documents receiving rank 1; since this is necessarily less than 50, we next include documents that receive rank 2. As soon as the number of different documents found in this way exceeds 50, we stop. The depth of this search is always at least rank 3, but we have seen cases in which it goes as low as rank 9. One consequence of this is that it clusters systems with respect to their performance on the most popular documents. As future work, one might instead consider clustering systems according to their performance on either the most popular relevant documents or even the most popular irrelevant documents. (The latter suggestion may seem strange, but it is arguable that common patterns of mistakes better identify similar systems than common patterns of correct retrieval.)

Regarding the block-clustering algorithm, our analysis used the routine 3M in BMDP (1985), with command lines:

```
/ PROBLEM TITLE='BLOCK CLUSTERING FOR TOPIC 336'.
/ INPUT VAR=24.
     FORMAT=FREE.
/ VARIABLE NAME=n, Brkly22,
     Brkly21, Brkly23, anu6alo1, anu6ash1, anu6min1, Cor6A1cl,
     Cor6A2qt, Cor6A3cl, att97ac, att97ae, att97as, CLAUG,
     CLREL, INQ401, INQ402, VrtyAH6a, VrtyAH6b, city6at,
     city6al, city6ad, LNaVrySh, LNmShort. LABEL=1.
```

```
/ GROUP    CUTPOINTS(2 TO 24) ARE 5,10,20,30,50
/ BLOCK NUMBER=5. REFINE.
/ END
```

BMDP is a commercially available suite of statistical programs. The 3M algorithm was developed by Hartigan (1975), and, because of combinatorial explosion, does not truly examine all possible permutations of rows and columns. Instead it assumes that the family of blocks forms a tree, the family of row clusters forms a tree, and the family of column clusters forms a tree. Under these assumptions, it proceeds iteratively; at each step, it finds the row-column arrangement that most reduces the deviation between the data and the tree model. This stepwise algorithm cannot guarantee any kind of optimality, but it has worked well in many applications, and the performance is insensitive to reasonable violations of the model assumptions.

A key point about Hartigan's algorithm is that it groups the values of the entries into a small set of categories. In most of our analyses, we used six categories, consisting of those ranks less than 10, those from 10 to 19, those from 20 to 29, and so forth, up to a final category of ranks that are at least 50. These cut-points were chosen to reflect a sense of the practical distinctions that typical users make in retrieval searches, but it is clear that other studies might select different groupings. Nonetheless, preliminary experimentation indicated that most results were not affected by reasonable changes in the cut-points.

In general, the results of these block analyses were disappointing. Often the algorithm made only small changes in the order in which the systems and documents were originally presented, indicating that it was finding no substantive structure. Figure 1 shows one of the best runs, done for Topic 336, on Black Bear Attacks. We suspect this topic showed more structure than was usual because there is no single keyword that systems could exploit. Thus systems had to search on combinations of keywords, creating clusters of systems that use chained keywords in similar ways.

To read figure 1, BMDP uses the convention that blocks are denoted by the symbols ' ' (a blank), '.', '−', '=', and '+', in order of decreasing block size. Blocks are rectangular arrays in the display, but they may contain elements that do not satisfy the criteria for block membership; in BMDP, such elements are shown by numerical values that give the deviation, but this creates visual clutter and so we replace the numbers by asterisks. Also, it is possible for the same block to appear in several non-contiguous pieces, and blocks can overlap other blocks—the BMDP convention is that the most recently formed blocks overlay those previously formed. It is common for the largest block to be the rectangular array formed by all or nearly all of the rows and columns.

For a block clustering to be successful, it is necessary that the total number of entries within blocks be a substantial fraction of the total number of entries. In figure 1, 73% of the entries are grouped into blocks. However, this large fraction is not sufficient to validate the clusters. One must also examine the interpretability of the clusters formed on the rows, columns, and entries to ensure useful structure has been found. In our application, we see that four of the five manual systems have been grouped together (Brkly23, anu6min1, CLAUG, and CLREL, but not LNmShort), that all three systems that search only on title are close (att97as, city6at, and LNaVrySh), and that five of the six systems that use the long narrative

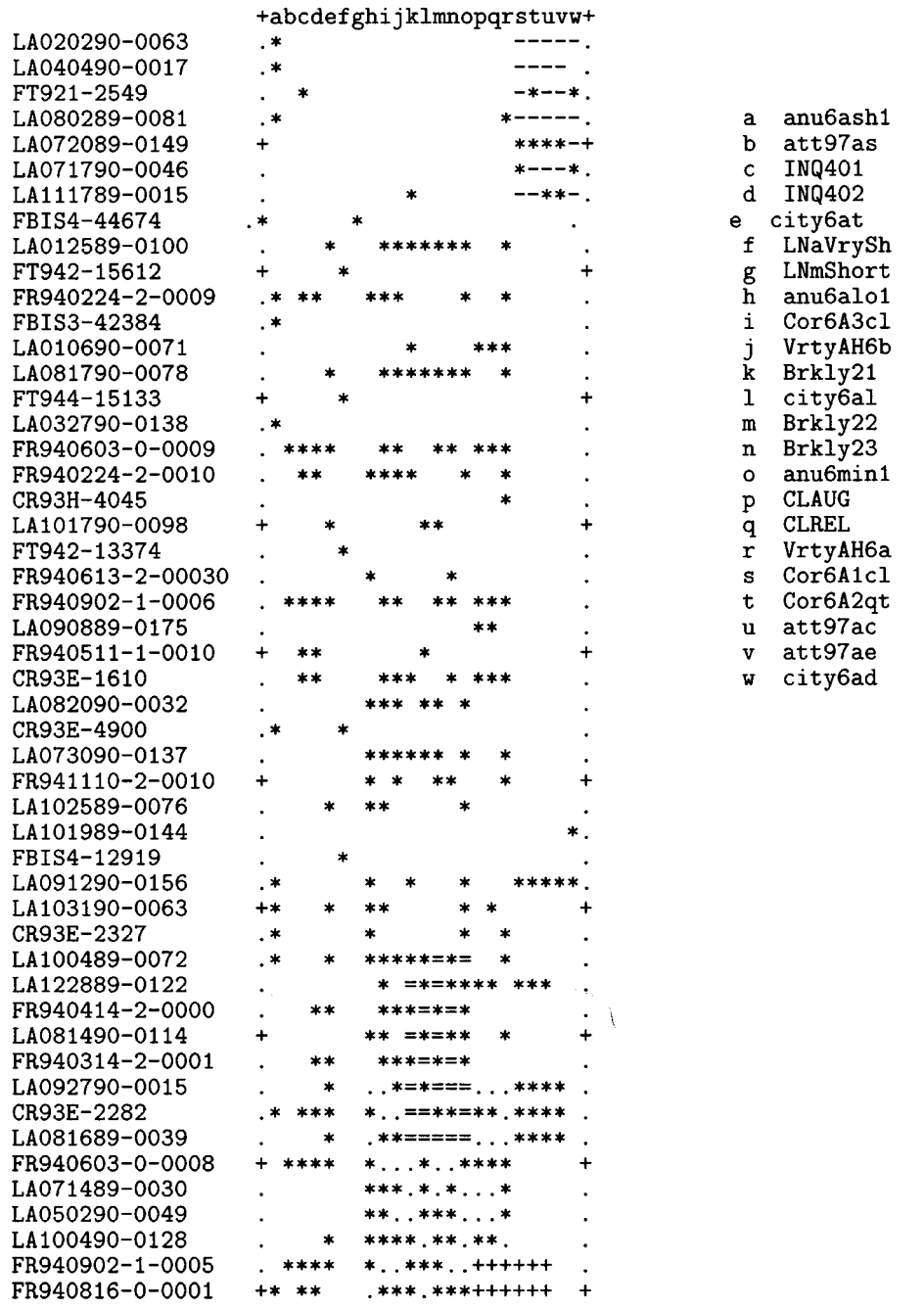


Figure 1. Block-clustering output for Topic 336.

are close (anu6alo1, Cor6A3cl, VrtyAH6b, city6al, and Brkly22, but not INQ402). Also, the algorithm puts the AT&T and Cornell systems together, as domain experts expect. We have not scrutinized the cluster structure among the documents, since that requires detailed information that we lack. Regarding the block structure within the matrix, this is very difficult to assess—some of the block divisions correspond to system clusters that appear valid (e.g., the division between Brkly22 and Brkly23 separates automatic and manual systems). But we don't wish to read too much post hoc explanation into this analysis.

The previous analysis performed block-clustering of systems and documents within Topic 336. We now look at block-clustering for systems and topics, where the performance of a system on a topic is measured by average precision. This analysis used the same 23 systems as before, and all 50 of the topics used in TREC-6. The new cutpoints are 0.9, 0.7, 0.5, 0.3, 0.1; these were chosen to capture qualitative performance differences. Figure 2 shows the results.

TREC experts feel that the cluster structure in figure 2 is slightly more interpretable than for figure 1, despite the fact that no penetrating new insight is obtained. First, we note that the number of entries that appear in blocks is 62%; this is slightly less than before, but may allow cleaner groupings. Second, the column clustering puts most of the Cornell and AT&T systems together (j-n), and correctly groups the title-only automatic searches (p-r) and the manual searches (a-c and v-w); other plausible structure is also apparent to those familiar with the history of search strategies of the retrieval systems. Third, among the rows, there is a tendency for the topics with the largest numbers of relevant documents to be grouped together (though this is not true for the topics with the fewest documents). Also, topics for which no clear keywords exist (e.g., 322, 327, 330, and 336) tend to be neighbors, and conversely (e.g., 302, 316, 326). But on balance the clustering among topics is less pronounced than that for systems.

It can happen that the block-clustering approach is too constraining, and that it would be better to cluster systems and topics separately. In principle, this is an inferior analysis, but given the mixed results from block clustering, we experimented with alternative cluster analyses.

There are many clustering algorithms—we use single-linkage clustering, since it sequentially combines clusters that minimize the distance between the closest objects in each (a nearest-neighbor clustering). This ensures that the analysis allows clusters to have peculiar shapes, rather than the ellipsoids enforced by most competing algorithms. When clustering systems, each system is represented by a point in $\mathbb{R}^{50}$, the coordinates of which are the average precision for that system on each of the 50 TREC-6 topics. Similarly, when clustering topics, each is a point in $\mathbb{R}^{23}$, the coordinates of which are the average precision from each of the 23 systems. Neither case seems likely to produce well-separated clusters with ellipsoidal contours, so single-linkage algorithm seems preferable. The code below for system clustering is from SAS (1996):

```
PROC CLUSTER METHOD=SIN;
VAR V1-V50;
ID SYSTEM;
PROC TREE HEIGHT=N SORT;
ID SYSTEM;
```

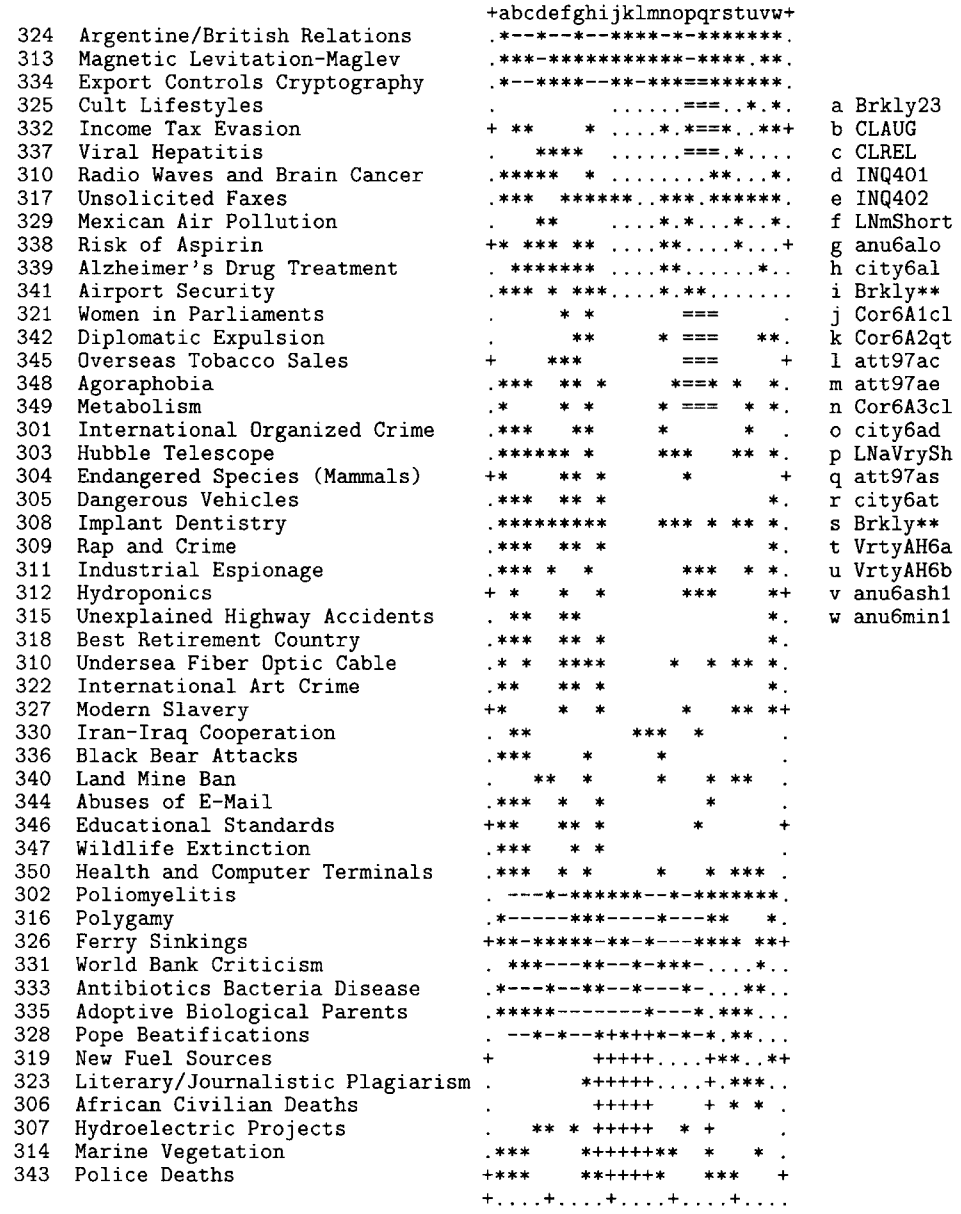


Figure 2. Block-clustering output for systems and topics.

Figure 3. Plot of SAS single-linkage cluster tree for 23 TREC-6 systems based upon average precision for each topic.

This performs a cluster analysis in which the similarity matrix is based on Euclidean distance between the systems in $\mathbb{R}^{50}$. There are many alternative methods and options that one could choose, but there is little domain knowledge to guide us in making the best selection.

The results for the system analysis are shown in figure 3. Experts see plausible groups in the early joining of att97ac, att97ae, Cor6a1cls, and Cor6A2qtcs, based on their development history. Similarly, two systems that use only the title (LNaVryShort and att97as), four that use the entire narrative (city6al, anu6alo1, Brkly22, and Cor6A3cll), and two that are manual (CLAUG and CLREL) form sensible groups. It is probably unsafe to attempt interpretation of later joins, as these are susceptible to chaining errors, a consequence of the nearest neighbor approach that results in spurious connections.

Similarly, the results from the topic analysis are shown in figure 4. Domain experts see much less interpretable structure here—although one can argue that some groupings are reasonable, this kind of judgment is too impressionistic to be useful.

Based on these explorations, our view is that this kind of cluster analysis is able to detect some structure in the data, but it does not produce significant new insights. The results are especially weak for topic clustering. Although one could undertake more exhaustive analyses that explore other cluster analyses and would likely obtain slightly more interpretable results, our preliminary studies do not warrant such effort.

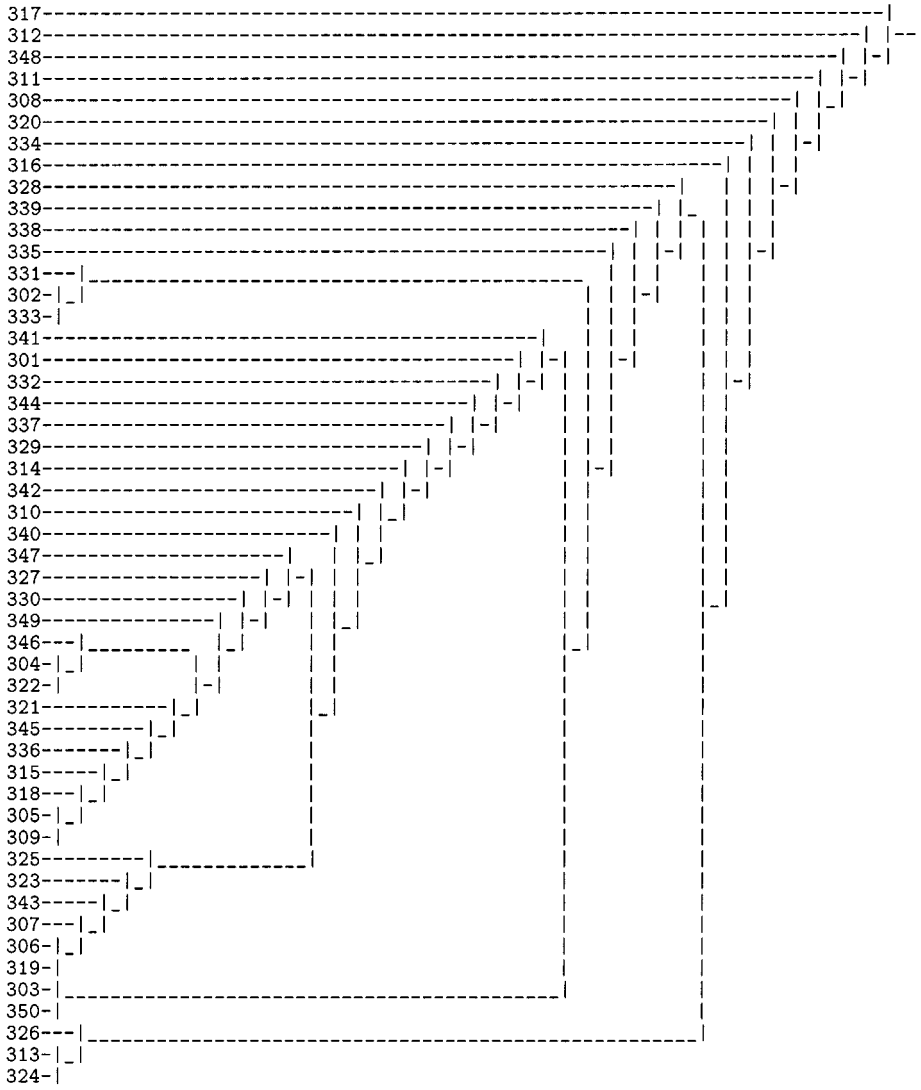


Figure 4. Plot of SAS single-linkage cluster tree for all TREC-6 topics based upon average precision for each system.

4. Rank correlations

None of the previous analyses has taken explicit account of the ranking information available in the retrieval orders. The natural approach is to calculate rank correlations between systems, using the ranks of the documents retrieved by each. To do this, the standard statistics must be altered to take account of the fact that not all documents are retrieved and ranked by both systems, and this entails some higher algebra.

All measures of nonparametric correlation are based upon some standardized metric, or distance, on the space of permutations. Let $\boldsymbol{u} = (u_1, \ldots, u_n)$ denote a permutation of the integers $1, \ldots, n$, where u_i is the permuted image of i. Then for two permutations $\boldsymbol{u}$, $\boldsymbol{u}'$, some standard metrics used in nonparametric correlation are:

$$
\begin{aligned}
d_1(\boldsymbol{u}, \boldsymbol{u}') &= \left[\sum_{i=1}^{n} (u_i - u_i')^2 \right]^{1/2} \\
d_2(\boldsymbol{u}, \boldsymbol{u}') &= \#\{(i, j) : u_i < u_j \text{ and } u_i' > u_j'\} \\
d_3(\boldsymbol{u}, \boldsymbol{u}') &= \#\{i : u_i \neq u_i'\} \\
d_4(\boldsymbol{u}, \boldsymbol{u}') &= \text{ number of transpositions needed to transform } \boldsymbol{u} \text{ into } \boldsymbol{u}'
\end{aligned}
$$

The first metric gives Spearman's rank correlation. The second gives Kendall's tau, and may also be interpreted as the minimum number of pairwise adjacent transpositions needed to transform $\boldsymbol{u}$ into $\boldsymbol{u}'$. The third gives Hamming distance—it counts the number of disagreements, and probably gives too little information. The fourth is Cayley's distance, which has convenient group-theoretic properties. Details on these and other metrics can be found in Diaconis (1988). For TREC applications, we follow Hull et al. (1997) in preferring Kendall's tau—it has a natural interpretation, and it enables one to calculate partial correlations for fully-ranked data.

The TREC document data is only partially ranked; i.e., each system ranks the top m out of n documents, and so the two lists need not include the same items. Thus one must modify the preceding metrics to define correlation measures over partial rankings. The key to this construction is the induced Hausdorff metric on the (right) coset space S_n/S_{n-m}. Essentially, the coset space is the set of equivalence classes obtained by restricting full rankings of n objects to partial rankings of m of the n objects, and the Hausdorff metric is the worst-case distance between the unobserved full rankings. By picking different distances, say from the list above, one can create different partial-ranking metrics. Critchlow (1985) provides a details, proofs, and computational algorithms.

To define the Hausdorff extension of the metric that supports Kendall's tau, let $\boldsymbol{v}, \boldsymbol{v}'$ be partial rankings of m out of n objects, where the ith component of the vector is an integer between 1 and n, inclusive, without repetition, for $i = 1, \ldots, m$. Take

$$
\begin{aligned}
A &= \{\text{objects ranked by both } \boldsymbol{v} \text{ and } \boldsymbol{v}'\} \\
B &= \{\text{objects ranked by } \boldsymbol{v} \text{ but not } \boldsymbol{v}'\} \\
C &= \{\text{objects ranked by } \boldsymbol{v}' \text{ but not } \boldsymbol{v}\}
\end{aligned}
$$

and set $h = \#B = \#C$. Then calculation shows that the corresponding metric on partial rankings is:

$$
\begin{aligned}
d_2^*(\boldsymbol{v}, \boldsymbol{v}') = \#\{&\text{pairs of items } (i, j) \in A \times A : v_i < v_j \text{ and } v_i' > v_j'\} \\
&+ h\left(n + m - \frac{h-1}{2}\right) - \sum_{i \in B} v_i - \sum_{i \in C} v_i'.
\end{aligned}
$$

Table 5. This matrix shows the distances, in terms of the Hausdorff extension of the Kendall's tau metric, between retrieval sequences of documents obtained from ten different systems for Topic 336, Black Bear Attacks.

System		1	2	3	4	5	6	7	8	9	10
1	Brkly23	0	975	1093	970	1094	1094	977	1096	854	850
2	Cor6A1cls		0	341	348	465	345	729	728	1103	732
3	Cor6A2qtcs			0	469	467	227	488	487	1216	615
4	Cor6A3cll				0	467	348	614	614	1096	728
5	att97ac					0	342	730	729	1101	733
6	att97ae						0	490	489	1101	496
7	CLAUG							0	223	1106	485
8	CLREL								0	1219	604
9	city6at									0	851
10	LNaVrySh										0

Measures of correlation are traditionally scaled to lie between -1 and 1, and this could be done here.

An example of such calculation is shown in Table 5, which gives the distances between ten of the retrieval systems for Topic 336, Black Bear Attacks. We chose this topic after experimentation with topics 326 and 341—these had less overlap among the top-ranked documents, possibly because there were larger numbers of relevant documents, and thus distances were larger and the results less illustrative. The systems were chosen because domain experts had prior beliefs about the outcomes. To avoid long computation, and to better mimic the kind of application real-world users typically make, we restricted attention to the pool of documents obtained by combining the top 40 retrievals from each of these ten systems; thus $n = 129$. For similar reasons, we took $m = 10$.

In interpreting Table 5, we note that experts correctly predicted that the Cor* systems would show relatively small differences, and that they would also be near the att* systems. But it was not anticipated that Cor6A2qtcs would be so near to CLAUG and CLREL. It was also surprising that the fully automatic short-title searches of LNaVrySh and city6at were not closer to each other, or to Cor* and att*, but were actually much like the manual searches of CLAUG and CLREL. This may reflect the fact that the title in this topic is a nearly complete list of the necessary keywords.

For statistical inference on the results in Table 5, one needs to know the distribution of this metric under the null hypothesis that the partial rankings are independent (i.e., have no agreement beyond that resulting from chance). But theoreticians have not yet been able to derive this, and so in practice one is forced to rely upon simulation to approximate the null sampling distribution. This poses no significant obstacles for TREC applications. For the choice of n and m above, a simulation of 10,000 random rankings finds that the mean distance between independent rankings is 1242.62, and the standard deviation is 97.2. Thus, for example, the distance between city6at and CLREL is not significantly different from chance, though all numbers less than 1048 would be significantly small (this assumes that a normal approximation is appropriate, which is reasonable, although the sampling distribution of the distances is skewed to the left).

There are two directions in which one might extend this approach. One is to the case of partial correlations, in which one looks for the association between two systems after controlling for the effect of a third—this could help in identifying how shared components (e.g., a thesaurus or stemming rule) of systems contribute to similar performance. The formula for partial correlation is

$$\tau_{ab\cdot c} = \frac{\tau_{ab} - \tau_{ac}\tau_{bc}}{\sqrt{(1-\tau_{ac}^2)(1-\tau_{bc}^2)}}$$

where τ_{ab} is the Kendall's tau correlation between systems a and b, and $\tau_{ab\cdot c}$ is the partial correlation between systems a and b after both have been adjusted for their agreement with system c. But it is difficult to make this work in the case of partially-ranked data, since the areas of agreement may not include overlapping sets of documents.

A second extension is to Mallows' model (1957), which enables a more flexible description of the distribution of ranks than the simple model of complete independence used in the simulation above. Computation is difficult, but the inferential strategy is a straightforward extension of a technique applied to graphs, trees, and partitions, and other combinatorial objects (cf. Banks and Constantine 1998). Let D_m be the space of all possible document sequences of length m. For a metric d on D_m, define the probability of observing sequence $\boldsymbol{v}$ as

$$p(\boldsymbol{v}) = c(\boldsymbol{v}^*, \sigma)\exp[-\sigma d(\boldsymbol{v}, \boldsymbol{v}^*)] \quad \forall \boldsymbol{v} \in D_m$$

where σ is a concentration parameter (analogous to the inverse of the variance), $\boldsymbol{v}^*$ is the modal sequence of partial rankings, and $c(\boldsymbol{v}^*, \sigma)$ is a normalizing factor. The key advantage of this model is that it has interpretable location and scale parameters, and thus many of the customary techniques of statistical inference are available. Also, the model allows users to choose the metric d so as to reflect a context-specific sense of distance.

Rank correlation methods seem a useful tool for TREC problems, but there are issues to resolve first. In particular, this analysis takes no account of the additional information on document relevance that is available. One might prefer a correlational analysis that restricted attention to just the subpool of relevant documents, or even to the subpool of irrelevant documents. Also, the area has unresolved theoretical problems concerning the distribution of the distances, the definition of partial correlations for partially-ranked data, and the calculation of maximum likelihood estimates for the Mallows model in this application.

5. Multidimensional scaling

Multidimensional scaling enables analysts to visualize distance-like data as a two-dimensional graph (sometimes higher dimensions are used). Given n objects $s_1, \ldots, s_n$ and a set of rough distances r_{ij} between each pair of objects, multidimensional scaling finds corresponding vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ in $\mathbb{R}^q$ such that the (usually) Euclidean distances $d(\boldsymbol{v}_i, \boldsymbol{v}_j)$ are as close as possible to the corresponding rough distances r_{ij}. The intuition is that sometimes

data has structure that can be captured in a simpler, low-dimensional representation. An introduction to this literature is in Young (1985).

The starting point of a multidimensional scaling analysis is a proximity matrix (also known as similarity matrix or dissimilarity matrix). Its entries are some measure of agreement or disagreement among the objects. One applies any of several variant algorithms to produce a plot whose points have interpoint distances that are approximately the same as the entries in the proximity matrix. A classic example is to take cities as the objects, and intercity driving times as the entries—then multidimensional scaling returns a plot in which the cities are relatively close to their true positions on a map (though the orientation of the map is arbitrary).

In the context of TREC data, multidimensional scaling has been used by Rorvig and Fitzpatrick (1998) on the documents in TREC-3; the basis for the proximity matrix was the number of search terms in common between pairs of documents. They found that document proximity is closely associated with relevance; i.e., where there are high densities of irrelevant documents, relevant documents are also found (which implies practical difficulties for search systems). However, this association may be an artifact of the pooling operation used to construct the set of possibly relevant documents. Rorvig et al. (1998) extend this analysis to visually identify the source database of the document. Together, these papers make a strong case that some document-level structure is found and visualized by two-dimensional scaling.

Our interest focuses more upon analysis of the retrieval systems. To this end, several possible proximity measures might be used; one could take $r_{ij} = 1 - |\rho_{ij}|$, where ρ_{ij} is Spearman's correlation between the ranks of relevant documents found by the ith and jth systems. Or one could ignore relevancy, and use a correlation measure based upon all documents, since common patterns of error can be better indicators of similar systems than patterns of ranking among correct retrievals.

Another approach, and the one shown in our example analysis, is to use symmetric set difference in the relevant documents retrieved by pairs of systems. Here one takes the sets D_i and D_j of relevant documents found by systems s_i and s_j, respectively, and calculates

$$r_{ij} = \#\{[D_i \cap (D_i \cap D_j)^c] \cup [D_j \cap (D_i \cap D_j)^c]\},$$

where superscript c denotes complementation. This is just the number of relevant documents found by exactly one of the two systems. If $r_{ij} = 0$, the systems agree completely on the relevant documents; if r_{ij} is as large as possible (i.e., equal to the number of relevant documents), then the two systems have disagreed on all of the relevant cases.

Symmetric difference is a metric, but not a Euclidean metric—the relationships among objects cannot be plotted on paper. Multidimensional scaling lets us discover whether the symmetric difference proximities show cluster structure or other interpretable patterns in the nearest approximate Euclidean representation. Different algorithms use different measures of discrepancy in finding the best approximation—we used PROC MDS from SAS (1996), set to minimize the ordinary least squares criterion in two dimensions. The command lines for the analysis of the proximity matrix built from topic 322 in TREC-6 are:

```
PROC MDS DATA=TOPIC322 LEVEL=ABSOLUTE OUT=OUT;
ID SYS;
OPTIONS PS=60;
PROC PLOT DATA=OUT VTOH=1.7;
PLOT DIM2 * DIM1 \$ SYS / HAXIS=BY 500 VAXIS=BY 500;
WHERE _TYPE_='CONFIG';
```

This code relies upon common defaults; much adjustment is possible, but is probably not warranted by the results shown below. The figures produced from this program were redrawn using Splus to improve graphical display.

One possible reason for the lack of interpretable structure is that two dimensions may be too few to represent the distance relationships among this data. To examine that, the MDS routine was run with the options changed to seek the best fit in dimensions 1, 2, 3, and 4. These results are shown in Table 6.

The badness-of-fit criterion must decrease monotonically as a function of the dimension. One inspects the criterion to discover whether there is a dimension for which a large drop appears, after which decline is relatively slow—should such a dimension exist, this suggests it is the correct dimension needed to approximately capture relationships implicit in the proximity matrix. From Table 6, Topic 322 does not seem to have a clear signature of this kind; there are large drops for both the two- and three-dimensional representations. But the topic is International Art Crime, which lacks specific keywords, and thus is relatively difficult. In contrast, both Topics 326 and 341 show a clear drop as one moves to two dimensions, with only small improvement thereafter. These topics (Ferry Sinkings and Airport Security) have relatively specific keywords, and are fairly easy.

Figures 5 and 6 give the plots produced by the multidimensional scaling algorithm for Topic 326, Ferry Sinkings, and Topic 341, Airport Security. In both cases, the change in badness-of-fit at two dimensions is particularly promising. Note that the two plots show considerable agreement in the relative positions of the systems (up to rotation, which carries no information in this analysis). In particular, note that systems 54, 60, 50, and 72, all occur on the convex hull of the scatterplot, and in the same order. Further, note the closeness of pairs 54 and 45, 55 and 59, 45 and 9, as well as many other consistent groups not labelled here (full labelling forces overstrikes). But there is also systematic distortion, as shown in the stretching of the separation between 54 and 60, and 55 and 9. Finally, some systems

Table 6. Values of the least squares badness-of-fit criterion obtained from SAS procedure MDS for three representative TREC-6 topics.

	Badness-of-fit		
Dimension	Topic 322	Topic 326	Topic 341
1	0.428	0.237	0.320
2	0.209	0.136	0.162
3	0.139	0.113	0.118
4	0.112	0.107	0.109

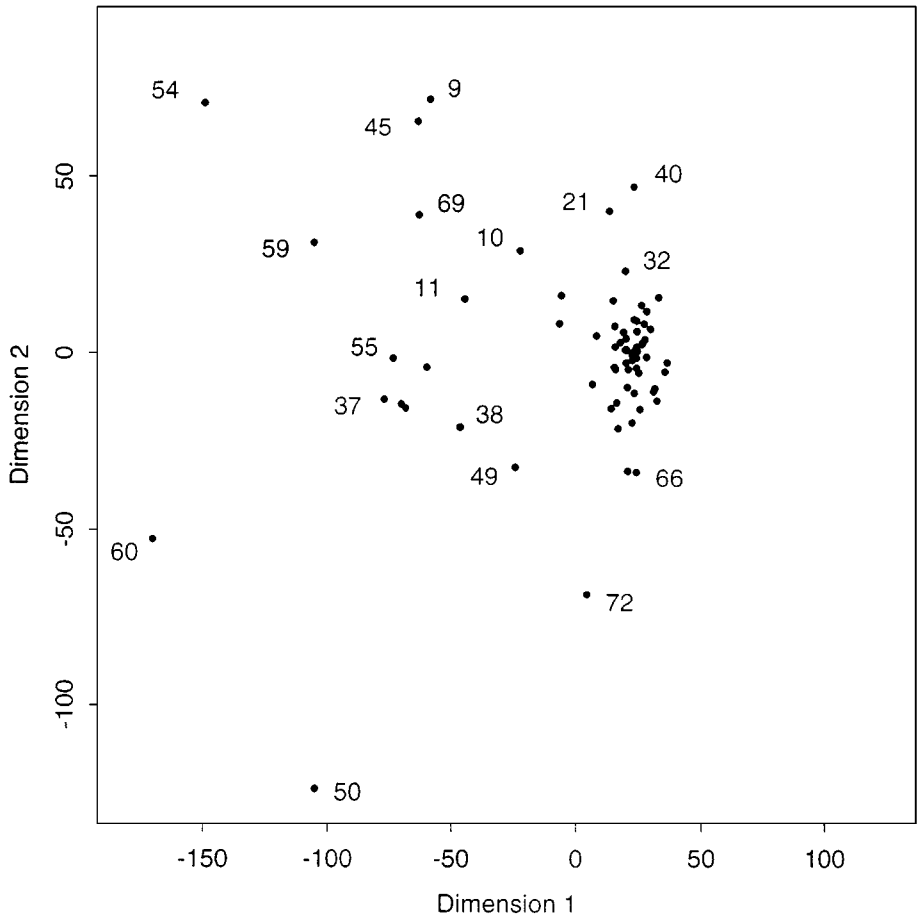


Figure 5. Two-dimensional scaling representation for Topic 326, Ferry Sinkings, for a proximity matrix based upon symmetric set differences among the retrieved documents.

show very different relative locations; for example, 65 moves from the outer boundary in Topic 341 to the interior for Topic 326, and there are many points between 37 and 72 in Topic 326, but not in Topic 341.

Besides assessing geometric agreement between figures 5 and 6, domain experts can also examine the plausibility of the neighbor relationships among these systems. Table 7 provides the numbering system for the retrieval systems shown in these figures. However, experts were not able to find sensible explanations for the patterns in the figures, not even in the case of the relatively conspicuous systems on the peripheries of the figures.

Overall, the multidimensional scaling approach suffers from several serious drawbacks. First, it does not seem to find strikingly useful patterns in the analyses we have done. Second, it is difficult to use when the badness-of-fit criterion suggests the correct approximating dimensionality is larger than two, or maybe three. Third, the results depend upon how

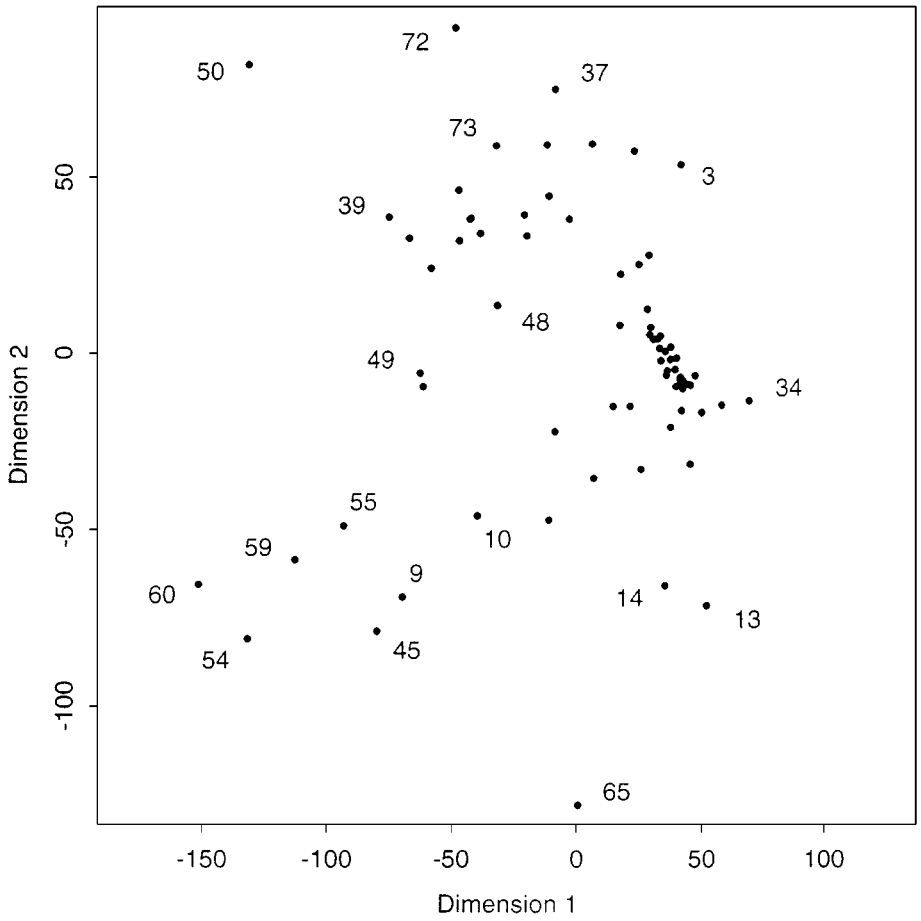


Figure 6. Two-dimensional scaling representation for Topic 341, Airport Security, for a proximity matrix based upon symmetric set differences among the retrieved documents.

the proximity matrix is calculated—we used symmetric set difference, but there are a very large number of other possibilities that one could try, and domain experts can provide little empirical basis for choosing one over another. Nonetheless, our pessimism does not guarantee this avenue is a dead end, but only that further progress requires extensive data churning, and even that may find nothing useful.

6. Beadplots

The similarity scoring in most retrieval systems are such that documents tend to be found in groups. Documents in the same group will have similar ranks for that system, but as a group the magnitude of their ranks can change across systems. For example, two systems might run the same search but with different weights on the keywords, so that the group containing

Table 7. Key to numerical codes for the information retrieval systems represented in figures 5 and 6.

1. Brkly21	2. Brkly22	3. Brkly23
4. CLAUG	5. CLREL	6. Cor6A1cls
7. Cor6A2qts	8. Cor6A3cl	9. DCU97lnt
10. DCU97lt	11. DCU97snt	12. DCU97vs
13. INQ401	14. INQ402	15. LNaShort
16. LNaVryShort	17. LNmShort	18. Mercure1
19. Mercure2	20. Mercure3	21. VrtyAH6a
22. VrtyAH6b	23. aiatA1	24. aiatB1
25. anu6alo1	26. anu6ash1	27. anu6min1
28. att97ac	29. att97ae	30. att97as
31. city6ad	32. city6al	33. city6at
34. csiro97a1	35. csiro97a2	36. csiro97a3
37. fsclt6	38. fsclt6r	39. fsclt6t
40. gerua1	41. gerua2	42. gerua3
43. glair61	44. glair62	45. glair64
46. gmu97au1	47. gmu97au2	48. gmu97ma1
49. gmu97ma2	50. harris1	51. ibmg97a
52. ibmg97b	53. ibms97a	54. ispa1
55. ispa2	56. iss97man	57. iss97s
58. iss97vs	59. jalbse	60. jalbse0
61. mds601	62. mds602	63. mds603
64. nmsu1	65. nmsu2	66. nsasg1
67. nsasg2	68. pirc7Aa	69. pirc7Ad
70. pirc7At	71. umcpa197	72. uwmt6a0
73. uwmt6a1	74. uwmt6a2	

one keyword will switch position with the group containing only the other keyword in the rankings.

To visualize this effect, we constructed beadplots. For a specific topic, the rows in a beadplot correspond to systems, and the “beads”, gray and colored diamonds, along each row represent documents. The position of a bead along a row indicates the rank in which the corresponding document was retrieved by the system associated with the row—beads to the left are ranked before beads to the right. The key point is that beads with the same color indicate the same document; the plot makes it relatively easy to spot documents that are retrieved together as a group, even when the ranks of those retrievals are very different. Such cases show up as splotches of the same color, at (possibly) different positions along the rows.

The colors assigned to the documents use spectral (ROYGBIV) coding. The ordering ranges from most relevant (dark red) to least relevant (light violet). Because there is no gold standard for ranking in TREC data, we initially used the ordering obtained from

the University of Waterloo's system, the "reference system", as this is widely believed to combine good performance with the human flexibility of a manual system. In subsequent analyses, a composite ranking was used instead; this composite combined information on the retrievals from all of the systems.

We found it useful to make two beadplots for each topic. The first shows the ranks assigned by each system to any relevant documents also retrieved at ranks 1–100 by the reference system, where relevance was determined by the topic proposer and the color coding was determined by the University of Waterloo rankings. Gray beads on this plot stand for any member of the set of relevant documents not retrieved among the top one hundred documents by the reference system. The second shows the ranks assigned by each system to any non-relevant documents also retrieved at ranks 1–100 by the reference system, where relevance was determined by the topic proposer and the color coding was determined by the University of Waterloo rankings. Gray beads on this plot stand for any member of the set of non-relevant documents not retrieved among the top one hundred documents by the reference system. The advantage of the second beadplot is that it can highlight documents or sets of documents that act as red herrings, fooling multiple systems in similar ways.

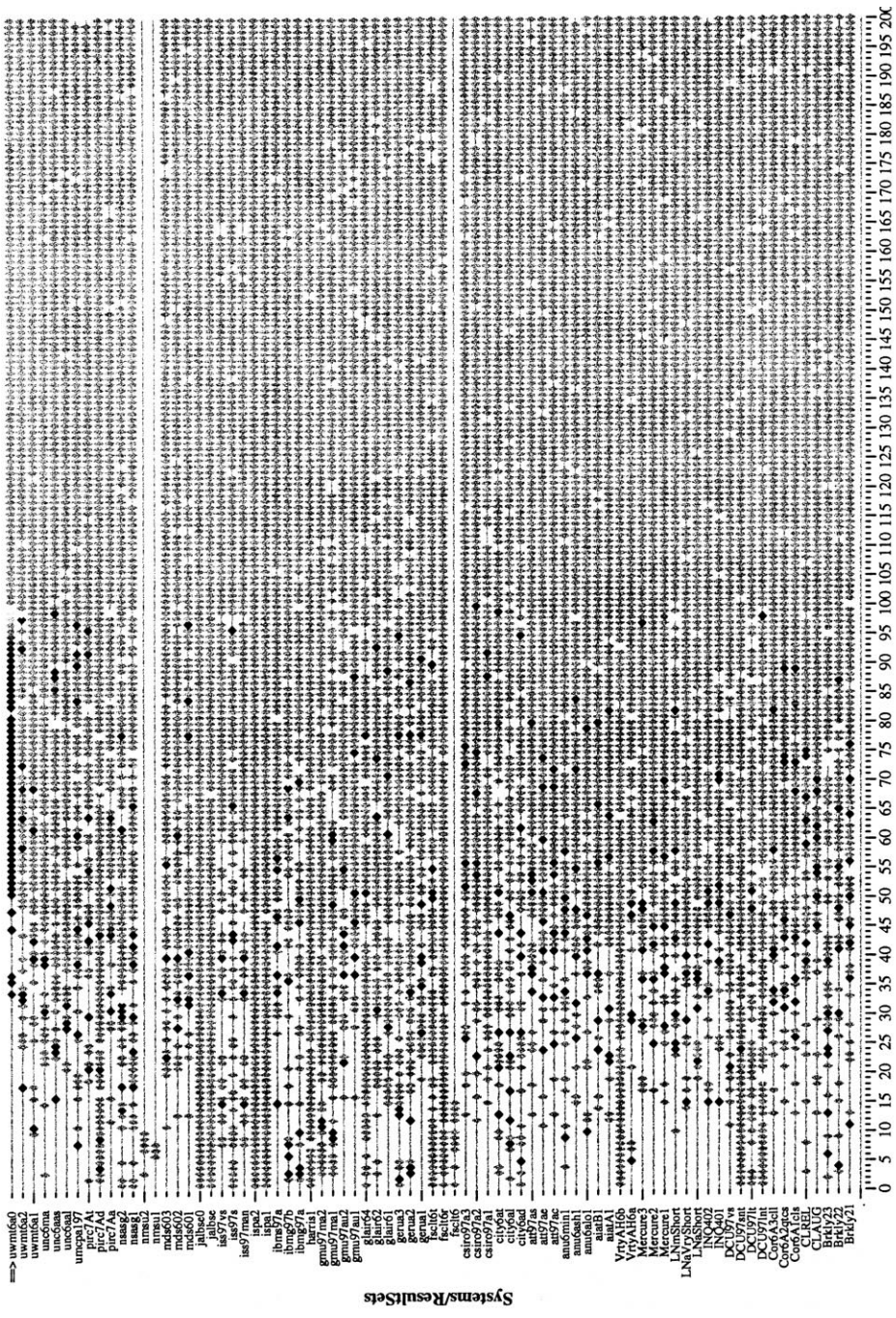
As an example of a beadplot pair, consider figures 7 and 8, which show the results for Topic 326, concerning ferry sinkings. Note that for figure 7, the beadplot of *relevant* documents:

- There is a tendency across nearly all systems for the colors to shift from red to blue as one moves left to right—this means most systems roughly agree with the University of Waterloo's rankings (shown in the third line from the top).
- Some sets of documents (e.g., yellow-oranges and light greens) tend to move together, either relatively forward or backward, as previously discussed.
- The density of the colored beads decreases as one moves to the right, indicating that most of the relevant documents are found early.

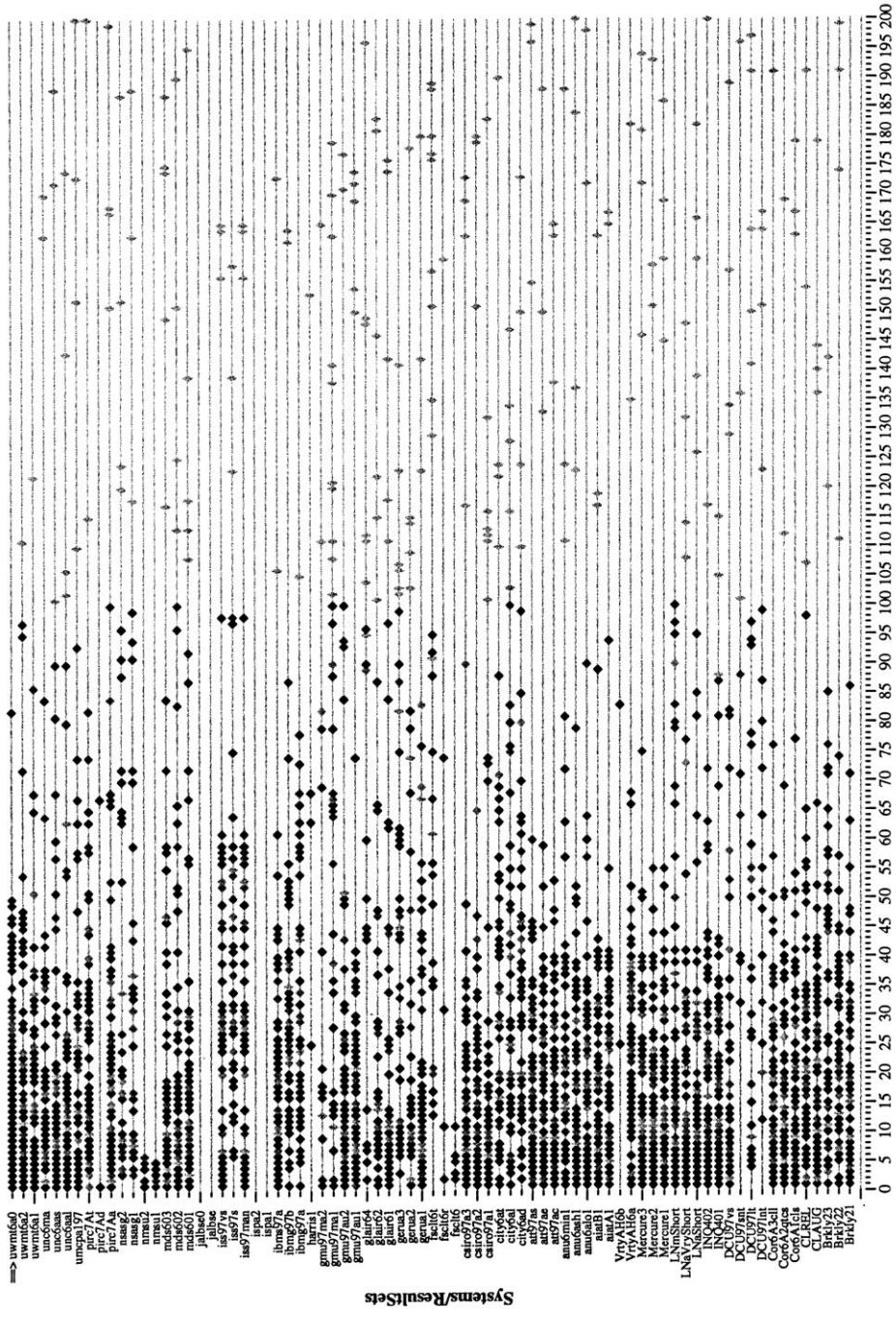
Similarly, for figure 8, the beadplot of *irrelevant* documents, one sees that:

- The color ordering is weaker, indicating less agreement on the rankings of irrelevant documents across systems.
- Some documents (green and blue) appear to be red herrings for many of the retrieval systems.
- The density of the beads is much less than before, and it decreases as one moves from left to right, showing that there is less agreement on retrieving the irrelevant documents, and that the amount of disagreement increases with retrieval depth.

Figure 7. Beadplots show the rank at which each relevant document was retrieved by each of the text-retrieval systems on topic 326. The rows correspond to the retrieval system, and the colored dots correspond to documents. Dots of the same color indicate the same document, and the order and spacings along the row indicate the ranks the documents were assigned.



Distribution of nonrelevant documents for topic 326 ranked 100 or higher by uwmt6a0



Distribution of relevant documents for topic 326 ranked 100 or higher by uwmt6a0

We have looked at all 50 beadplot pairs from TREC-6, and find that there is useful visual variety between the topics. One can identify at a glance topics for which there are many red herrings, or few relevant documents, or marked differences in performance across systems. For some topics, the absence or presence of documents in the University of Waterloo set and/or the composite ranking set is conspicuous, implying either that the reference ranking missed important documents, or that groups of systems failed to find important documents, respectively.

The visual features in the beadplots have led us to attempt to place the topics in a three-dimensional space. One dimension is a measure of how diverse the systems are with respect to performance on the topic; another is a measure of how easy the topic is, in terms of the mean average precision (i.e., the average across the topics of the average precision on each topic) of the best systems; the third is a measure of the number of red herrings. Currently we are exploring whether the representation of topics in this space makes sense to domain experts and points up interpretable structure among the topics.

7. Item response analysis

Item response analysis is used in educational testing. The simplest model is the Rasch model, which assumes that the ith student has ability level $\alpha_i > 0$ and that the jth question has difficulty $\lambda_j > 0$. Then the probability that student i answers question j is modelled as

$$p_{ij} = \frac{\alpha_i/\lambda_j}{1 + \alpha_i/\lambda_j}$$

and the parameters can be estimated if one has a large number of students all taking the same set of test questions.

In the context of information retrieval, the analogy is that the ith document retrieval system has an ability level $\alpha_i > 0$, and that the jth document has difficulty level λ_j. But the Rasch model is too simplistic for this application. The key deficiencies are:

1. There are two types of error a document retrieval system can make—it can recover an irrelevant document, or it can miss a relevant document. A good model should capture both kinds distinctly.
2. The Rasch model does not allow for the chance of a correct guess (if $\alpha_i = 0$ or $\lambda_j = \infty$, then $p_{ij} = 0$). But in document retrieval, a system could guess about relevance and be correct a significant fraction of the time.
3. The probability of correct classification, as a function of ability level, changes more rapidly for some topics than others.

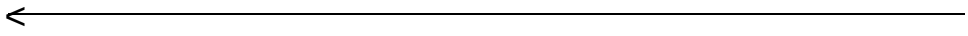


Figure 8. Beadplots show the rank at which each irrelevant document was retrieved by each of the text-retrieval systems on topic 326. The rows correspond to the retrieval system, and the colored dots correspond to documents. Dots of the same color indicate the same document, and the order and spacings along the row indicate the ranks the documents were assigned.

To a degree, these problems can be addressed by employing more flexible models. However, this requires extensive computation and the conclusions typically become sensitive to the validity of the modelling assumptions.

Educational testing research has developed models that are more pertinent to the needs of information retrieval scientists, especially in the context of multiple choice questions where guessing is an effective strategy. The three-parameter logistic model is an example; here

$$p_{ij} = c_j + (1 - c_j)\frac{\exp[Da_j(\alpha_i - \lambda_j)]}{1 + \exp[Da_j(\alpha_i - \lambda_j)]} \tag{1}$$

where α_i and λ_j are interpreted as in the Rasch model, D is typically set to 1.7 so that the curve closely resembles a Gaussian cumulative distribution function, c_j is the probability that one can obtain the correct answer by guessing, and a_j is the discrimination parameter, which controls how rapidly the curve rises as a function of ability (cf. Kolen and Brennan 1995). Software to fit this model is commercially available; BILOG 3 (Mislevy and Bock 1990) is probably the most sophisticated.

This model is better, but still does not capture some of the issues in document retrieval. For example, one might consider two separate models, both of the form in (1). The first model would provide the probability of correctly identifying a relevant document, and second the probability of correctly identifying an irrelevant document. This doubles the number of parameters, and leads to two kinds of ability parameters. These two parameters, plotted against each other, would be a natural representation of the capability of the system, analogous to a point on an ROC (Response Operating Characteristic) curve.

The problem with this kind of item response analysis is that it is probably not a sufficiently good model to describe the performance of information retrieval systems. Such systems typically assess relevance according to a large number of unequally-weighted document features. If two systems may differ in only their weighting of one feature, they can have widely dissimilar performances for a topic in which many of the documents exhibit that feature. The problem is akin to one that arises in educational testing, where a question can draw upon both mathematical and verbal skills; the standard solution in educational testing is to develop latent trait models that allow multivariate ability parameters. But the classical statistical models for this, based upon the multivariate normal distribution, appear inappropriate for complex feature-weighting systems.

Based on previous results, especially comovement patterns seen in the beadplots, and on knowledge of the kinds of weightings used in many retrieval systems, we believe that some extension of latent trait models may be possible. This extension must have one ability dimension for each feature, and then reduce the total dimensionality according to duplication in the features' coverage. But this would entail substantial micro-modelling, and is probably not a usefully practical solution.

8. Conclusions

Our purpose in these analyses was to discover whether there were real differences among the retrieval systems, and to understand how such differences relate to differences among topics and collections of documents. None of the work we have done using the six approaches

discussed here has provided the sort of insights we were seeking, and the prospects that additional work along these lines will yield significantly better results vary but are not generally promising.

The search for simple additive models in the analysis of variance seems for all practical purposes closed, but there is some hope that careful models of interaction might be useful. Significant extensions of theory and its application are required in the case of rank correlations and item response analysis. Better use of existing means—better choice of clustering algorithm, better visualization of clustering results, alternate proximity matrices as a basis for multidimensional scaling, use of measures other than average precision—might improve the picture, but we can provide no evidence to support this hope.

We suspect that we will get further by looking at future data in a more narrow setting. This will involve shifting our focus from aggregate measures, such as average precision, to richer descriptions of IR system behaviour. This shift might also entail use of

1. full lists of ranked documents, such as those visualized by the bead plots,
2. analyses that incorporate more information on system design, as opposed to the black-box approach taken in this paper,
3. designed experiments that select sets of similar topics to provide pseudo-replicates of retrieval performance,
4. carefully framed hypothesis tests, designed to probe the effects of small changes in the retrieval systems.

However, the fundamental finding of this paper is that this elephant is hard to understand. Conventional approaches do not seem to work, and we should be prepared to recognize that sometimes no simple conclusion can account for the observed complexity.

Acknowledgments

We thank the Edison National Historic Site for their assistance. Additionally, we benefited from useful discussions with Donna Harman, Ellen Voorhees, and Walter Liggett as well as from the comments of the anonymous reviewers.

References

Banks DL and Constantine GM (1998) Metric models for random graphs. Journal of Classification, Vol. 15, 199–224.

Critchlow DE (1985) Metric Methods for Analyzing Partially Ranked Data. Springer-Verlag, New York, NY. David HA (1981) Order Statistics. Wiley, New York.

Diaconis P (1988) Group Representations in Probability and Statistics. IMS Lecture Note Series, Vol. 11. Institute of Mathematical Statistics, Hayward, CA.

Dixon WJ, ed. (1985) BMDP Statistical Software: 1985 Printing. University of California Press, Berkeley, CA.

Dyer FL and Martin TC (1910) Edison: His Life and Inventions. Harper and Brothers, New York.

Harman DK (1994) Overview of the second text REtrieval conference (TREC-2). In: Harman DK, ed. The Second Text REtrieval Conference (TREC-2). U.S. Government Printing Office, Washington, DC, pp. 1–20.

Harman D, ed. (1996) The Fourth Text REtrieval Conference (TREC-4). National Institute of Standards and Technology Special Publication 500-236. U.S. Government Printing Office, Washington, DC.

Hartigan JA (1975). Clustering Algorithms. Wiley, New York.
Hull DA, Kantor PB and Ng KB (1997) Advanced approaches to the statistical analysis of TREC information retrieval experiments. Report presented at the TREC-6 Conference at the National Institute of Standards and Technology.
Kolen MJ and Brennan RL (1995) Test Equating: Methods and practices. Springer-Verlag, New York.
Lawrence S and Giles CL (1998) Searching the world wide web. Science, 280:98–100.
Mallows C (1957) Non-null ranking models I. Biometrika, 44:114–130.
Milliken GA and Johnson DE (1991) Analysis of Messy Data, Volume 2: Nonreplicated Experiments. Van Nostrand Reinhold, New York.
Mislevy RJ and Bock RD (1990) BILOG 3: Item Analysis and Binary Scoring with Binary Logistic Variables, 2nd ed. Scientific Software, Mooresville, IN.
Rorvig M and Fitzpatrick S (1998) Visualization and scaling of TREC topic document sets. Journal of Information Processing and Management, Vol. 34, 135–149.
Rorvig M, Sullivan T and Oyarce G (1998) A visualization case study of feature vector and stemmer effects on TREC topic-document subsets. Proceedings of the 1998 Annual Meeting of the American Society for Information Science, Information Access in the Global Information Economy, CM Preston, Ed. Vol. 35, 130–142.
SAS Institute, Inc. (1996) SAS/STAT Software: Changes and enhancements through release 6.11. SAS Institute, Cary, NC.
Tague-Sutcliffe J and Blustein J (1995) A statistical analysis of the TREC-3 data. In: Harman DK, ed. Overview of the Third Text REtrieval Conference (TREC-3). U.S. Government Printing Office, Washington, DC, pp. 385–398.
Young FW (1985) Multidimensional scaling. In: Kotz S, Johnson NL and Read CB, eds. Encyclopedia of Statistical Sciences. Wiley, New York, Vol. 5, pp. 649–659.