Pergamon

Information Processing & Management Vol. 34, No. 2/3, pp. 135–149, 1998

Printed in Great Britain
0306-4573/98 $19.00 + 0.00

PII: S0306-4573(97)00058-7

VISUALIZATION AND SCALING OF TREC TOPIC DOCUMENT SETS

MARK RORVIG[1]* and STEVEN FITZPATRICK[2]

[1]School of Library and Information Sciences, University of North Texas, P.O. Box 311068, Denton, TX 76203-6796, U.S.A.

[2]Center for Measurement and Evaluation, 2616 Wichita, P.O. Box 7246, University of Texas at Austin, Austin, TX 78713-78246, U.S.A.

(Received 11 June 1997; accepted 19 September 1997)

Abstract—Although numerous new document visualization tools are emerging throughout academia and industry, reliable test data for such tools has not yet been established. This paper explores the applicability of the TREC Information Retrieval Test Collection for this purpose using commonly available data and statistical methods. © 1998 Elsevier Science Ltd. All rights reserved

1. INTRODUCTION

In the Summer of 1996 at the Second Annual Workshop on Visual Information Retrieval Interfaces (VIRIs) held in conjunction with the Association for Computing Machinery Special Interest Group on Information Retrieval Annual Conference, the topic of evaluation of VIRIs occupied the center of discussion (Rorvig & Hemmje, 1998). One of the dimensions of this discussion was the potential for the use of the TREC/Tipster (Harman, 1993, 1994, 1995, 1996) data collection as a means of testing VIRIs. This topic was not new, and followed the presentation of two works at the conference which had explored aspects of this issue. The first (Hearst & Pedersen, 1996) described the use of TREC data clustered by vector cosine methods and visualized by cluster centroid topics. The second (Veerasamy & Belkin, 1996) used TREC data to evaluate the ability of users to improve search strategies in an interactive environment in which the output of searches by relevance was visualized. However, what had not been done previously was the exploration of TREC documents through visualization and scaling, to learn what native characteristics this collection might possess when analyzed through such tools.

Although the TREC collection has had a very significant impact on the field of IR, little is known about the behavior of TREC Topic–Document sets in a visual field. Moreover, since there are many methods of visualization of retrieved results, it is unclear how one should proceed. There is a great temptation to proceed with a known system. However, while proceeding from a known system may illuminate various aspects of the system under study, it is certain that such use will not illuminate the test collection itself. Whatever the strengths, weaknesses, and biases of a given system might be, they can act only as confounding variables on the understanding of the TREC data. Therefore this study uses only general techniques, common to a variety of data analysis tasks. No technique has been used which is not readily available and accessible.

This exploration is timely due to the emergence of marketplace tools for VIRIs (Text Navigator, ThemeMedia). Additionally, outside this marketplace, there are perhaps as many as 50 different systems which are under trial or development (Korfhage *et al.*,

*To whom all correspondence should be addressed. Tel.: 940-565-2445; Fax: 940-565-3101; e-mail: mrovig@jove.unt.edu.

1995; Gupta & Jain, 1997). However, the benefits of VIRIs (which rely on a host of methods for data aggregation) over and above that of retrieval systems which simply list documents by their estimated degree of relevance, is, in the main, anecdotal. Workers in the field of VIRI development badly need a measurement tool. This study examines the characteristics of ten TREC Topic–Document sets to explore the potential for use of TREC data for this purpose.

2. PROCEDURES

TREC data were obtained for this study by purchase of Volume I of the TREC three-volume document collection of the Linguistic Data Consortium of the University of Pennsylvania. This volume contains source documents for five subsets of TREC: Associated Press (AP) wire feeds; Department of Energy (DOE) documents; Federal Register (FR) documents; Wall Street Journal (WSJ) full texts; and sources from Ziff-Davis Publishing. This last set was not used for this study because there was no correspondence between the Volume I Ziff documents and the relevance judgments rendered on these documents for the first ten topics. The topic files and records of relevance judgments were obtained from the National Institute of Standards and Technology FTP site maintained for researchers wishing to work with TREC data and its various subsets.

A decision was made early on to use only documents from TREC which had actually been judged by a qualified human for relevance to a topic. This decision also logically entails certain assumptions about the underlying linguistic density of these datasets, since they were compiled through the use of the 'pooling method' (Sparck Jones & Van Rijsbergen, 1975) in which TREC-1 participating system retrieval postings (top 100 documents for each query) narrowed the potential documents required for human judgment from 3300 to 1279 (Harman, 1996). Thus, a number of documents which might have produced lexical scatter in the visual field were excluded from consideration by the methods of this study.

TREC Topic–Document sets were culled from the collection for the first ten topics. These ranged in size from a low of 407 documents for Topic 10, to a high of 680 for Topic 2. For each topic, a joint probability matrix of document relationships was created using the basic coefficient measure

$$\mathrm{sim}(\mathrm{doc}_i,\mathrm{doc}_j) = \frac{\left[\sum(\mathrm{term}_{ik} \cdot \mathrm{term}_{jk})\right]}{\sum \mathrm{term}_{ik} + \Sigma \mathrm{term}_{jk}} \tag{1}$$

where document distance is computed as the sum of the number of unique terms (k) which two documents (i,j) share divided by the total number of unique terms between them for all k in i and j. This coefficient takes values in the range of 0.0 to 1.0. There are at least five other classic measures which could have been used, most common among them the Dice and Jaccard measures (see the discussion in Salton & McGill, 1981, pp. 201–204 for further study). However, again in the absence of prior knowledge of the effect of these measures on the document distance score, parsimony dictated use of the simplest possible measure of eqn (1). Further, no attempt was made to stem words, or remove stopwords, though different measures of document similarity would be expected had these methods been employed. Finally, no weighting scheme was applied to either words or documents, since nothing is known about the impact of these techniques on visualization and would again have introduced new factors for evaluation.

The joint probability matrix for each topic was then scaled using the Multidimensional Scaling (MDS) procedure in SAS (SAS Institute Inc., 1996). MDS was chosen for this analysis of document distances since it is, if not the most robust measure for scaling, at least one of the most widely understood and enjoys a rich

research tradition within the field of information science (for examples, see: Katter, 1967; Weis & Katter, 1967; Katter *et al.*, 1971; Small, 1973; White & Griffith, 1981; Rorvig, 1988; McCain, 1990; Rorvig *et al.*, 1993; Larson, 1996; Goodrum, 1997). There are other techniques which might have been used (Rorvig & Hemmje, 1998), but each of them in turn would have introduced new dimensions which would have to have been considered. All of these alternatives are grounds for further investigation and research.

Background material and comprehensive bibliographies for the relationship between MDS, scaling, and the field of psychometrics in general may be found in Rorvig (1988). Goodrum (1997) updates the literature in the area during the period between the two publications.

In this context, it is also important to distinguish scaling from clustering. Although MDS procedures result in aggregates of documents when visualized as plots (which shall for convenience of discussion be referred to as clusters in this paper), the goal is not to cluster like items but rather to locate documents in a Euclidian space such that the distances in the solution space are related to the inter-document similarities computed from eqn. (1). Clustering is the assignment of like objects to a common set by criteria established under control of the researcher. As Jain points out (Dubes & Jain, 1979), anything can be clustered but not everything can be scaled. In VIRI systems, when clustering is used, clusters are represented typically by abstractions (for example circles), rather than dots in a metrically defined space, as they are in this paper.

While both MDS and clustering methods use proximities data as input, hierarchical clustering methods locate clusters in a discrete space of high dimensionality. There are as many dichotomous dimensions as there are nodes in a tree diagram representing the clustering solution. MDS in contrast, uses a continuous, usually Euclidian, space of low dimensionality to locate the objects being scaled. The proximities among a set of N objects can be perfectly represented in an MDS solution of dimensionality $N-1$, but the data analyst usually wants to adequately represent the similarities among the objects in a space of much lower dimensionality. Also, in hierarchical cluster analysis, objects can belong to only one cluster at any level of the tree. MDS locates objects in a solution space such that the distances between the objects in the solution space are related to the proximities among the objects in the data. The distances between clusters in the discrete space of cluster analysis often cannot be related to the proximities data by a linear or even monotone, function (Davidson, 1983). MDS and cluster analysis are often used as complimentary methods. One can scale objects using MDS and then use the distances from the solution space as input to cluster analysis.

Formally, the strictest case of the classic MDS model may be expressed as

$$\delta_{i,j} = d_{i,j} = \left[\sum (x_{i,r} \ldots x_{j,r})^2\right]^{1/2} \tag{2}$$

where $\delta_{i,j}$ is the dissimilarity (also referred to as the proximity) between objects i and j, $d_{i,j}$ is the distance between objects i and j in the solution space, and $x_{i,r}$ and $x_{j,r}$ are the coordinates obtained directly from them. Such a simplistic model will rarely suffice because the proximities data are typically human judgments or other derived measures about the similarity of objects which contain error. Furthermore, the proximities may be measured on an ordinal, interval or ratio level and it may not be appropriate to treat them directly as distances.

An early modification of the MDS model, known as nonmetric MDS, fitted an unspecified monotonic transformation of the distances to the proximities. Other modifications involve transforming the proximities with a power or logarithmic function, and/or transforming the distances from the model. This development is quite important in the context of this paper, since as demonstrated below, it is indeed the power transformation which yields the most coherent visual organization of the TREC datasets under study. Moreover, as suggested in Rorvig (1998), this organization appears to have strong consequences for retrieval performance as well.

Additionally, observers who may be unfamiliar with MDS are often puzzled by the lack of axis labels and axis scale marks. Such labels are not provided for MDS because the scale is recovered from the document interpoint similarities, and their derived interpoint distances, and not from the identification of continuous variables along an axis.

The first attempt at scaling the documents of Topic 1 ($n = 586$) under the assumption that the level of document distances is ordinal yielded most dissatisfying results as shown in Fig. 1. The scattering of relevant documents throughout the collection yields the conclusion that only poor retrieval would be possible since this dispersion violates the proximity hypothesis for these data. The proximity hypothesis stipulates that like documents should aggregate in like spaces (Van Rijsbergen, 1989; Voorhees, 1985). Similar treatments of the data at absolute value, interval, and ratio level measurement assumptions yielded similarly disappointing results: relevant documents tended to scatter at random throughout the collection of documents for each topic.

Finally, a maximum likelihood estimation (MLE) procedure which uses a power transformation was tried. The specific technique, initially developed by Ramsay (1977) and implemented through the Newton–Raphson algorithm (Hendry, 1995), is designed to be used when considerable skewness or kurtosis is suspected in the distribution of item similarities. In this case, specifically, some documents are highly related to one another, but only weakly related to other documents overall. Moreover, it is also used when the similarities are assumed to contain a large error term, perhaps arising in this study as a result of the parsimonious similarity measure used to produce the initial matrix of similarity coefficients. Table 1 provides a table of the power exponents and slopes for all ten TREC Topics, together with a 'badness' of fit estimate rendered through the power transformation technique.

The results of this procedure, i.e. loglinear in the SAS command set, when applied to these data were considerably more rewarding as shown in Fig. 2 below. Ramsay proposed a maximum likelihood procedure for obtaining the stimulus coordinates of the MDS solution. His reasoning was twofold. First, proximities data are positive and the lognormal distribution has a constant ratio of the standard deviation to the mean. Ramsey observed that objects which enjoy high similarity coefficients will also enjoy a sampling distribution of low variability. Objects which are dissimilar will have a sampling distribution which is more variable. Specifying an error model for the data allows the use of maximum likelihood for estimation of the solution space. The lognormal distribution is shown below as

$$P(\delta|d,\sigma^2) = (2\pi)^{1/2}(\sigma\delta)^{-1}\exp\left[\frac{-\ln(\delta|d)}{2\sigma^2}\right] \tag{3}$$

where δ is an observed dissimilarity rating, d is the true distance between the two objects, and σ is the standard deviation. With this error function, the log likelihood is

$$\ln L = \sum\sum \ln P(\delta_{i,j}|d_{i,j},\sigma^2). \tag{4}$$

Thus, given the assumption of dissimilarity among various members of the TREC subcollections, it was desirable to test the kurtotic expectations for data submitted to this technique by examining the degree to which the algorithm organized dissimilar document types as reflected in the various collection subsets of this Topic. Figure 3 replots this result, with each collection subset displayed as a separate color.

Figure 3 reveals that document types do, in fact, exert a strong influence on the overall shape of these document locations. This appears to be especially strong for elements of the subset of Federal Register (FR) documents, one cluster of which is completely isolated (but also contains no relevant items.) That Federal Register data are significantly different from other data types is noted elsewhere. Harman (1996, p. 7) commenting on the causes of the increased difficulty of the routing task in TREC-4

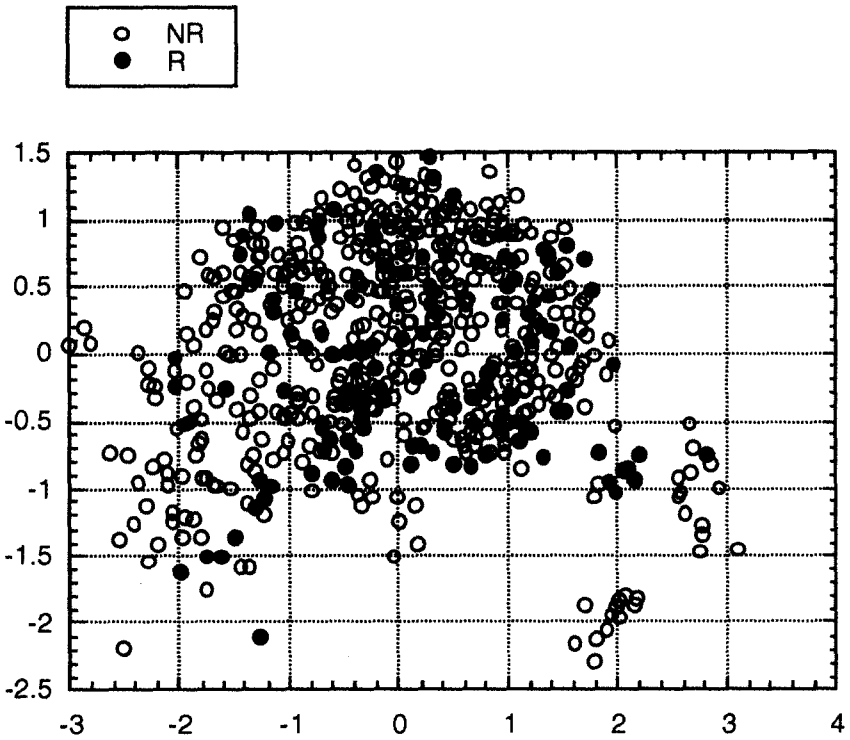

Fig. 1. TREC Topic 1 documents, ordinally scaled by MDS (n = 586).

attributes this in part to "...the concentration of long Federal Register documents, which have consistently been harder to retrieve..."

Overall, however, these results are positive in terms of the ability to use TREC data for VIRI testing. If, for example, a user were to probe at the darkest areas of this plotted shape, some relevant documents would be found immediately. This feature of correspondence between dense inter-document structures and relevant document presence appears in nine out of the ten topics explored in this study.

Finally, it should be pointed out that the blue dot in Figs 1–3 lying at the coordinates of $(-2.60, -2.50)$ is actually the Topic document itself. The Topic documents for the first 50 Topics of the TREC collection are quite long and consist of detailed descriptions of the criteria to be used by a judge in assessing relevance. In creating the joint probability matrix for the first five topics, the Topic document was simply included among the other documents. The expectation was that since the Topic document is related in some way to all the relevant documents, it should therefore behave by plotting among them. This did not prove to be the case. In every instance, the Topic document distances from the central coalescence of relevant documents is very great. Moreover, this condition persists across all scaling treatments. This observation receives further attention in Section 4 of this paper.

Table 1. Overall fit, slope and power measures for all ten joint probability matrices. The number of iterations to convergence is variable among these data, with four failing to converge even after 100 iterations (*)

Topic no.	Badness of fit	Slope	Power	Number of iterations to convergence
1	0.11	0.73	0.14	74
2	0.14	0.68	0.17	79
3	0.12	0.72	0.15	100*
4	0.12	0.68	0.18	90
5	0.12	0.70	0.16	100*
6	0.13	0.68	0.16	58
7	0.12	0.71	0.15	67
8	0.13	0.68	0.16	57
9	0.13	0.70	0.13	100*
10	0.12	0.71	0.13	100*

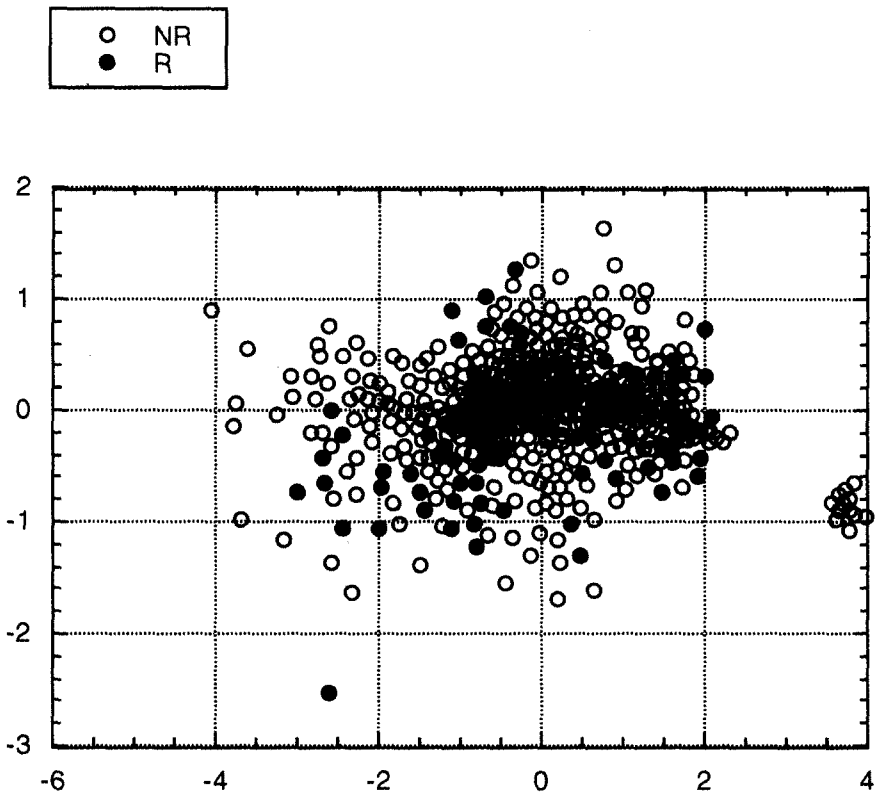

Fig. 2. TREC Topic 1 derived by a power transformation method (n = 586).

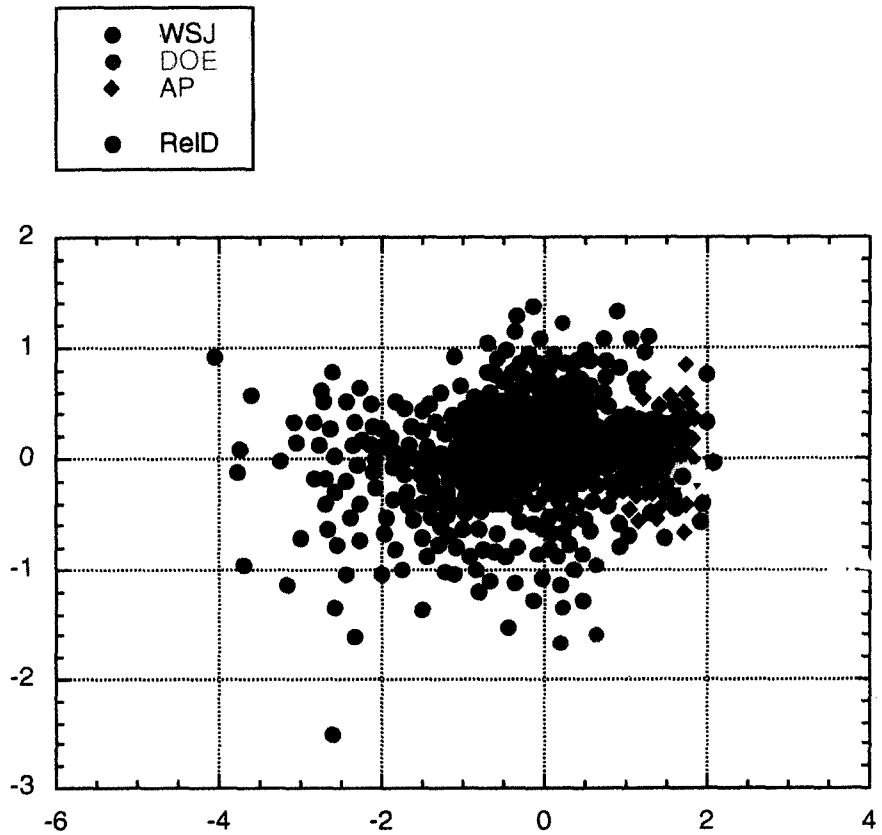

Fig. 3. TREC Topic 1 plotted by document type. RelD indicates the original relevant document set plotted throughout the sub-collections (n = 586).

3. RESULTS

Table 1 sets forth the pertinent evaluation measures for Figs 2 and 4–12 constituting the first ten topics of the TREC IR test collection. Table 1 yields very uniform measures of slope, power and fit. This suggests that the reliability of this technique in rendering uniform treatments of these data with respect to varying topics is satisfactory for further experiment and investigation.

Plot 5 for Topic 3 yields another interesting finding. Whenever the relevant documents do not coalesce into a definite center, or, as in the case of Topics 9 and 10, bifurcate into several centers, the MLE fails to converge within 100 iterations. This phenomenon may be observed in Figs 5, 7, 11 and 12 for Topics 3, 5, 9, and 10. Since there are, in all these figures, definite aggregates of non-relevant documents, it would appear that the greater the dispersion of relevant documents, the slower the solution convergence rate. This observation in turn leads to another possibility: visualization techniques may be useful to TREC experimenters who conduct failure analysis on queries. Since it is reasonable to expect that the broader the dispersion of relevant documents, the less lexical coherence exists among them, it follows as a consequence that it would be more difficult to retrieve them without suffering the penalty of low precision scores.

4. DISCUSSION

The results of this study yield a rich set of insights. Firstly, TREC data, when examined by the assumptions of skewed distributions and high error terms, behave in a remarkably orderly manner. Relevant documents aggregate consistently, and document type influences on their distribution can be readily observed in isolation from the dispersion and aggregation of relevant items. This finding establishes the applicability of these data for testing VIRI metrics and metaphors. This result is critical because, in VIRI expositions, it is quite easy to introduce artificial data relationships (Korfhage, 1991). Use of a standard collection would allow VIRI designers to test for these anomalies and suppress or mitigate them.

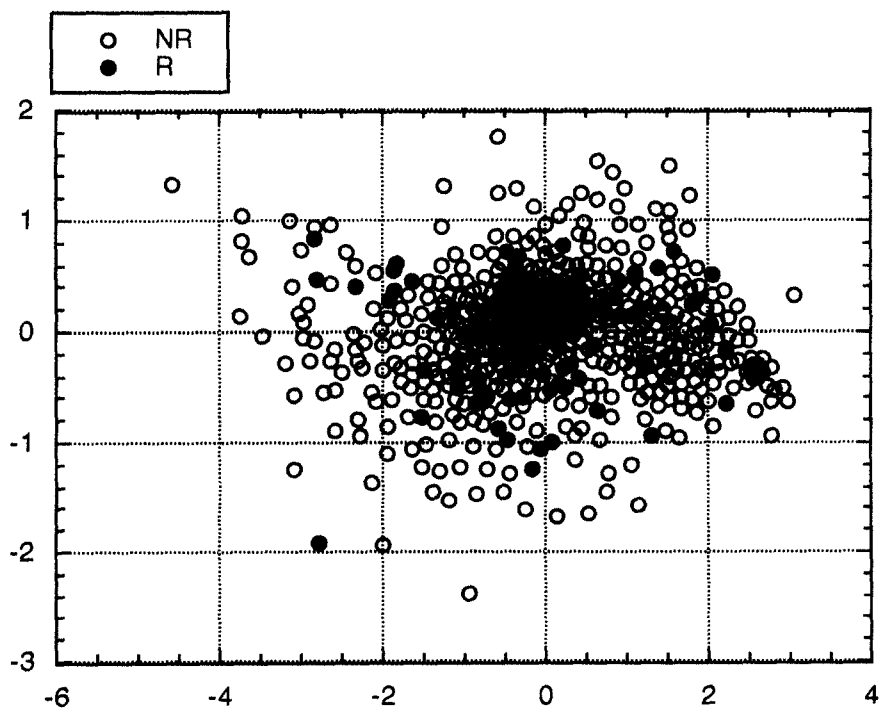

Fig. 4. Plot of Topic 2 document locations (n = 681).

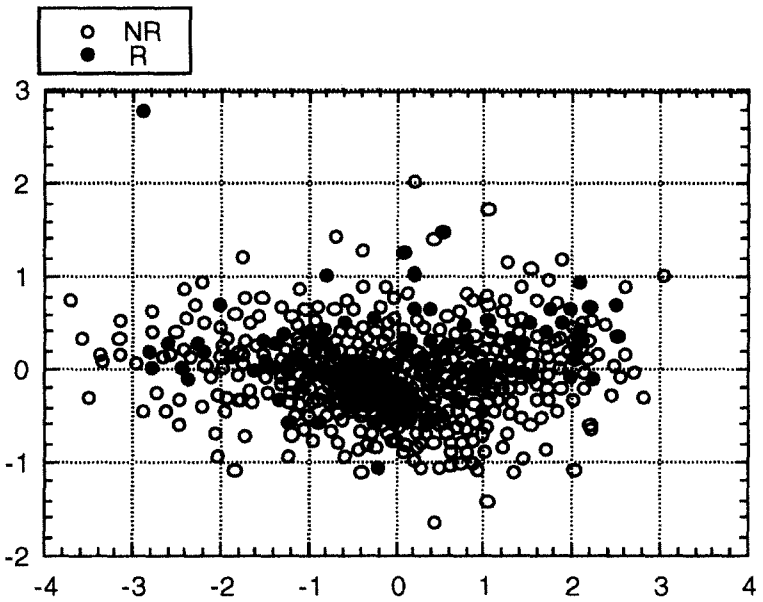

Fig. 5. Plot of Topic 3 document locations (n = 531).

This does not, of course, imply that TREC is the only such test collection which could be used for this purpose. Rather, it suggests that other test collections (for example those noted in Shaw *et al.*, 1997) should also be studied by visual scaling by a variety of similarity measures so that some means of comparing the fit of test collections with VIRI analysis might be obtained.

Second, for nine out of ten of these topic datasets, Topic 8 yielding the only major exception, it would appear that high inter-document proximity is a virtual proxy for relevance. Wherever there exist high densities of non-relevant documents, relevant documents are also found.

The significance of this finding is, however, mitigated by the prior selection method used to obtain these data in the first place. The pooling method treatment noted earlier implies that significantly greater lexical coherence exists among the data in these

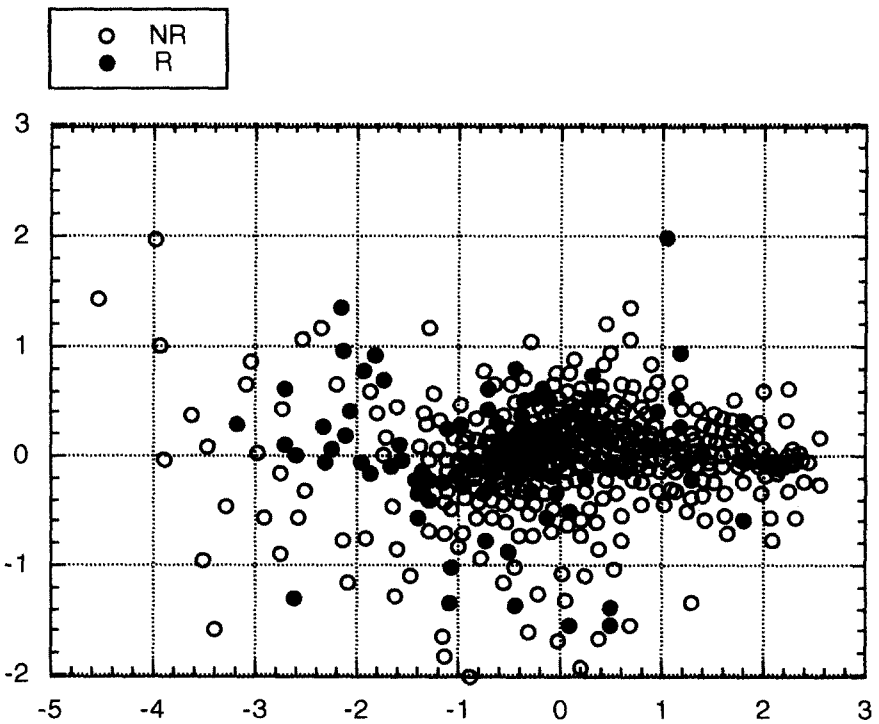

Fig. 6. Plot of Topic 4 document locations (n = 464).

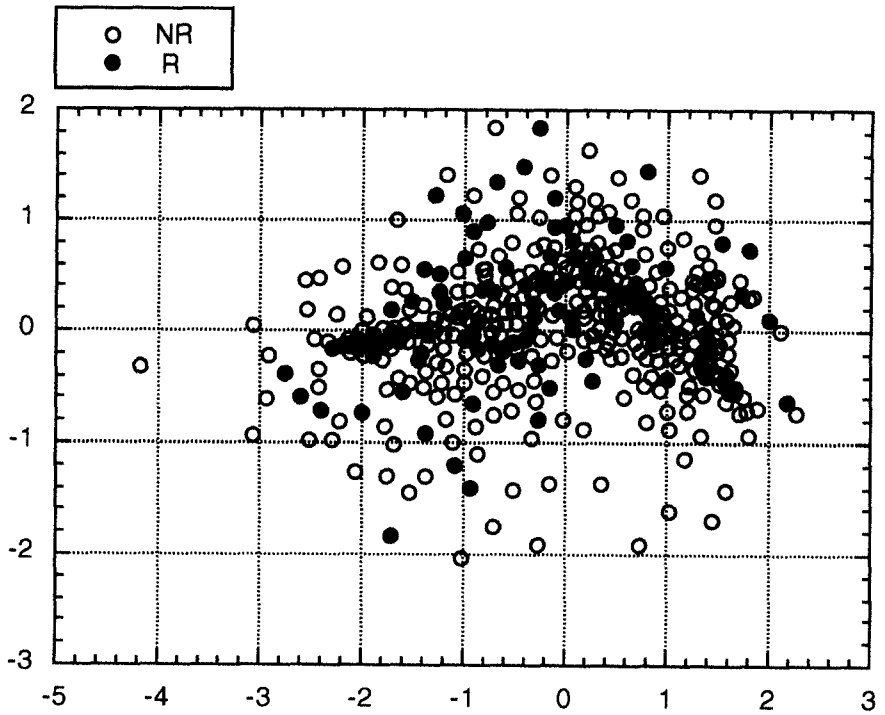

Fig. 7. Plot of Topic 5 document locations (n = 464).

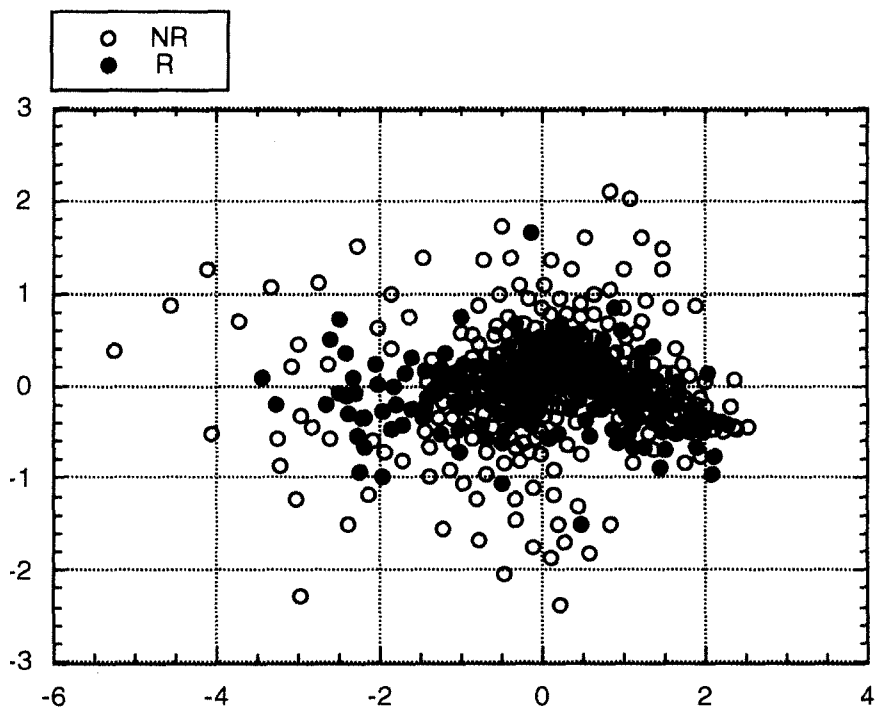

Fig. 8. Plot of Topic 6 document locations (n = 480).

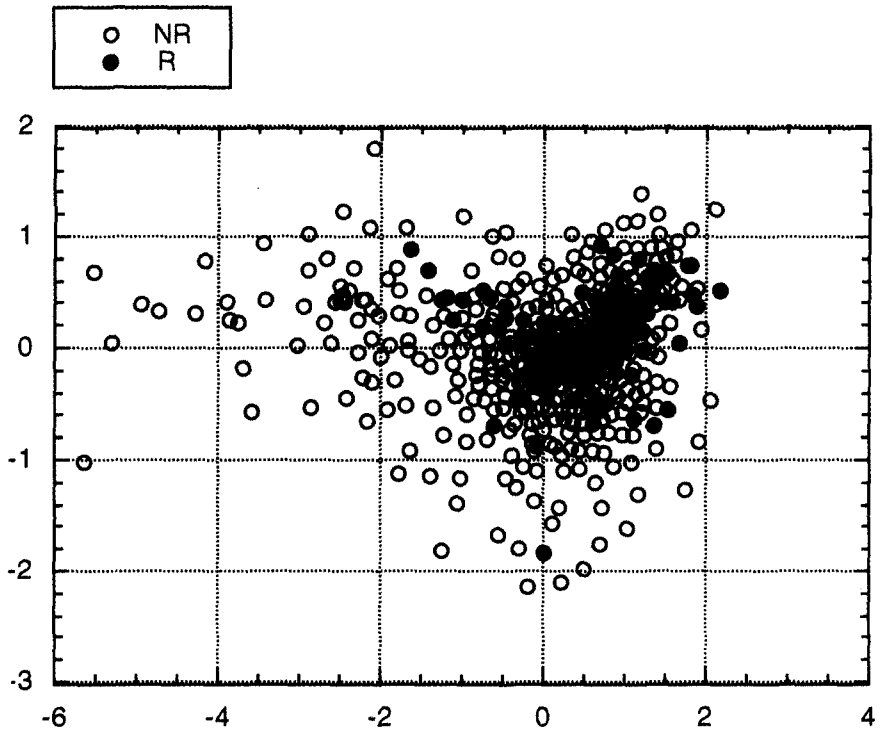

Fig. 9. Plot of Topic 7 document locations (n = 491).

datasets than might otherwise be obtained from a random search of WWW-based materials, for example. Nevertheless, the observation of relevant document appearance within high inter-document density structures does causally address the generally good performance of the many clustering techniques which have been used in IR; that is, the higher the inter-document proximity, the better the retrieval performance. Moreover, when this is not true, as in the plot of Topic 8, this phenomenon would work strongly against high retrieval performance in cluster retrieval systems. Inter-document proximity would be high for large regions of the dataset which contained few, if any relevant documents and lower precision scores would result from pursuit of any retrieval strategy which retrieved highly interrelated documents. Criticisms of cluster retrieval methods may be caused more by this dispersal phenomenon than from some general failure of cluster methodologies and their related assumptions as claimed by Shaw *et al.* (1997).

However, when a low density of relevant documents occurs, or inter-document densities do not signal the presence of relevant documents, as in Topic 8, lower precision scores would inevitably result in cluster-based or partial-match systems. There would be no general lexical relationship for the system to seize upon as a clue to the retrieval of relevant documents. In this instance, Boolean searches which examine documents for the presence of particular terms would remain the most effective form of retrieval since, in this case, it is 100% more certain that a document which contained a particular term would be more relevant than one which did not, no matter what its lexical relationship to other documents through other terms. A possible area of inquiry is the review of problematic TREC topics which yield low precision scores by examination the inter-document distance conditions which may lead to these scores.

The inter-document density observation also suggests a further empirical rationale for the generally good performance of one of the standard methods of search in partial match systems (i.e. a relevant document is identified by a searcher or search intermediary and resubmitted as a query). The increased density of the search pattern identified as relevant would simply permit the system to focus on the most dense correlates of the reformulated query. Although finding the first relevant document might be difficult, once found and resubmitted, results would improve because a dense

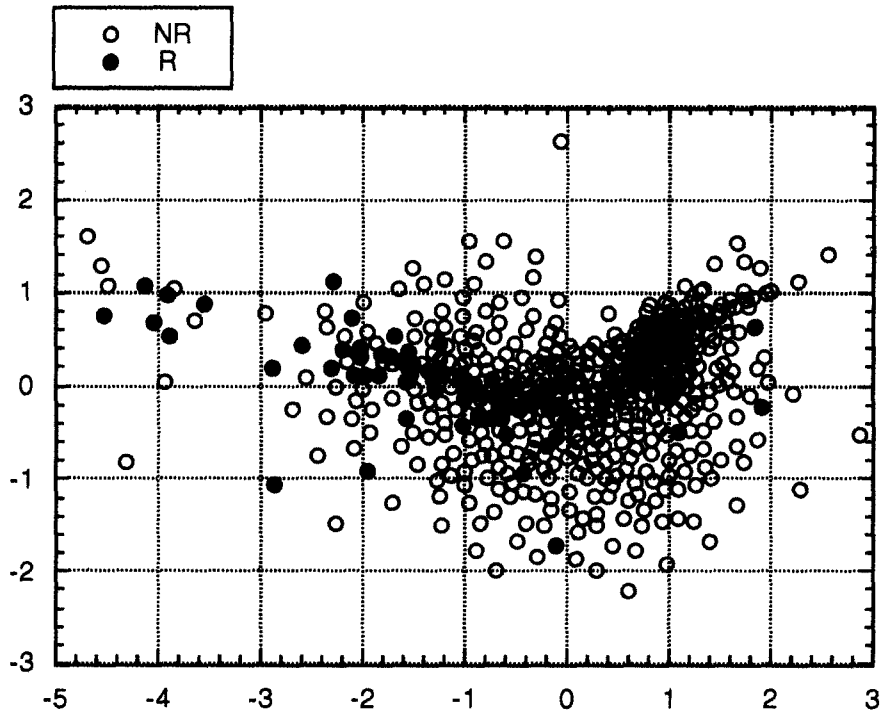

Fig. 10. Plot of Topic 8 document locations (n = 587).

inter-document proximity region could then be identified by the system. There is also some possibility that query–document feedback methods might be optimized by choosing for feedback not only documents judged relevant, but specifically a relevant document located within a dense inter-document location structure (Rorvig, 1998).

On the negative side, it should be carefully noted that the orderly behavior of these datasets comes at the price of very high computational overhead. Each of the joint probability matrices submitted to SAS for this study required from 1.5 to 2.5 h CPU time to complete on a Sun Ultra Enterprise 5000 class machine running at 167 MHz with 4 CPUs and 1 Gb of RAM. Presently, this technique is not suitable for interactive use.

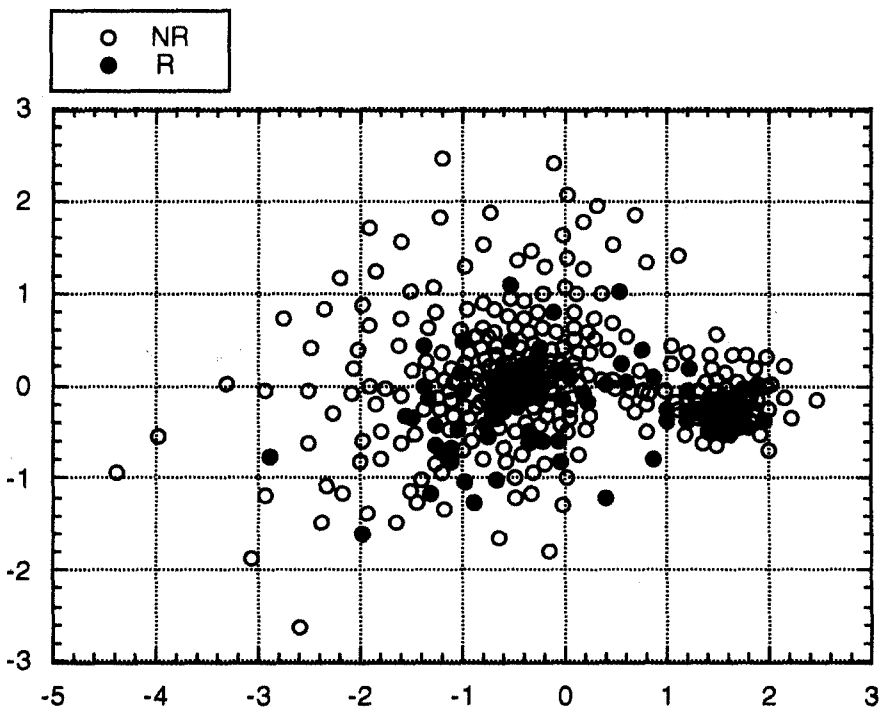

Fig. 11. Plot of Topic 9 document locations (n = 421).

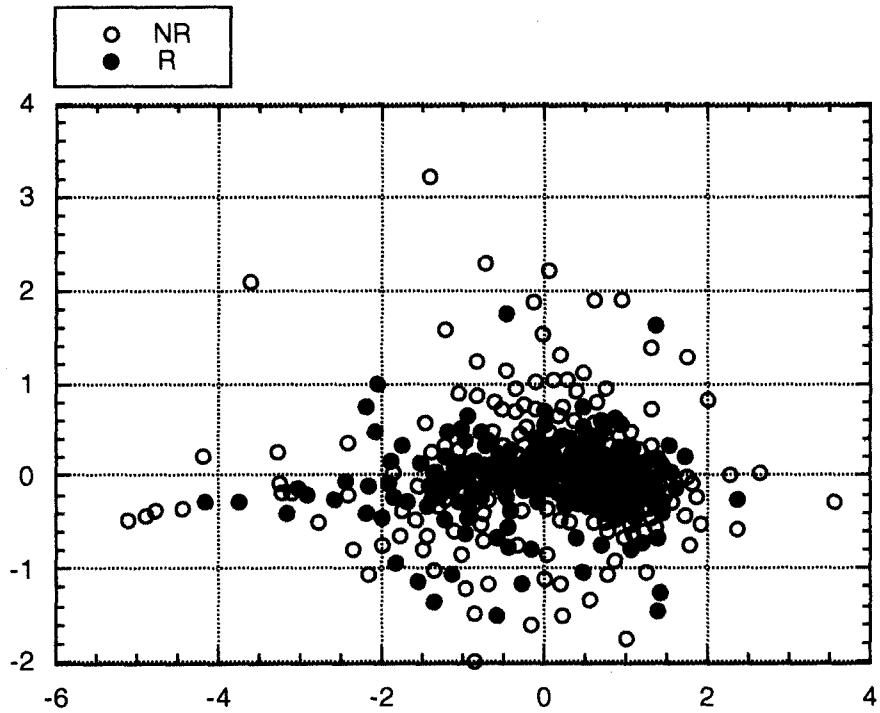

Fig. 12. Plot of Topic 10 document locations (n = 407).

The relationship between the choice of the original similarity measure and the efficiency of downstream processing of the resulting matrix of similarity coefficients deserves further study. Although it might well be that the loglinear MLE technique reflects innate properties and unique requirements of these data, it could certainly be hoped to be otherwise a manipulable variable of the choice of similarity measure. An optimal measure for this process would be one which maximally reduced the error term generated between the initial similarity measures and their transformed distances.

At large also is the issue of the relationship of the Topic documents to the other documents in the set. For Topics 1–5 for which this relationship was examined, Topic specification documents plotted at positions, (−2.60, −2.50), (−2.79, −1.89), (−2.89, −2.80), (−2.62, −1.29), and (−2.00, −0.74), respectively. In every case, these positions are quite far from the center of the visual field. Moreover, as noted earlier, this effect was observed regardless of the procedure used to scale the data. Since the Topic documents are related to all the relevant documents, the expected relationship was that topics should plot among the relevant items. Further study is necessary to assess this result, however, and such studies could readily be created from the available briefer queries used in the TREC-3 experiments in which the topic specification sections of the query statements were excluded.

A further source for error in the topic statement distance finding is the very simple measure of similarity used in this study. It may be the case that this simplest of measures fails to capture the relationship between the topic documents and relevant documents of the same set. A Dice, Jaccard, or even Cosine Vector similarity measure might yield significantly different relationships. However, since the relevant documents aggregate closely in conformity with expected proximities, this alternative hypothesis remains speculative.

In any case, whatever it is that the TREC relevance judges were doing, it is no doubt true that one thing they were not doing was counting the number of term tokens which various documents shared with the topic specification. Human beings bring to the judgment process a wealth of prior experience, synonym memory, and mental models of document relatedness and aboutness (Buckland & Gey, 1994). Therefore it should not be surprising that since judges evaluate documents on conceptual rather than lexical correspondence, topic specification documents would not necessarily be found at close

proximity to topic documents. This study's finding suggests that creation of proxy specification documents which include many synonyms would be necessary to 'put the topic specification document in its place' so to speak. Indeed, some TREC-4 experiments do focus on query expansion (see for example, Satoh *et al.*, 1996).

Related to this phenomenon of topic–document distance is the potential for study of the degree of paramorphism between the ranked relevance output of text search systems and the plotted document locations. What one would like to see in these rankings in terms of a paramorphic hypothesis is the appearance of a number of highly ranked items within the closely aggregated relevant items. This result would be expected because highly ranked items are expected to be closely related to relevant items. Concurrently, the reverse should be true for items scoring near the bottom of the ranked retrieval list: documents of little relevance should be those at the fringes of the plot aggregates. The required alteration of the topic specifications in terms of synonym additions in order to achieve such paramorphism may shed light on operational requirements for fully automatic query procedures, since the more centrally located the query within the visual field, the better its anticipated retrieval performance.

Finally, these results, particularly for Topic 8, suggest that for certain topics, there may simply be unsurpassable limits to retrieval performance. It may be possible to alter the query, or change the strategy from partial match to Boolean (or vice-versa), but the results may always be less than desirable. Consider the phenomenon of dispersion in Bradford's law, for example. Given a discipline which is immature, the dispersion rate for key documents among journals may be extremely high. This case is due to the discipline itself, and not to any failure of Bradford's law. Similarly, for certain topics and the document sets which are associated with them, no amount of tinkering with any component of the retrieval process may yield improvement in terms of precision and recall. Visualizing these Topic–Document sets at least presents this problem in light of the objective criteria of joint probability similarities and their transformed document distances, rather than test results reported without reference to an underlying degree of dispersal.

There are over 200 TREC queries, and this study has examined only ten of them. Patterns which are suggestive for this small sample may not hold over larger document sets. However, the possibility for the use of this technique in failure diagnosis of VIRI and traditional IR systems is apparently warranted.

5. CONCLUSIONS

This study reveals the suitability for TREC datasets as a development tool for VIRI interfaces. The study also suggests that visualization techniques of TREC may be useful in the evaluation of patterns of performance with respect to individual topics by experimental text retrieval systems. Of particular interest is the observation from nine of the ten datasets that a high degree of inter-document proximity appears to be closely related to the presence of relevant documents. The computational cost of organizing these data into useful visual fields is high and may be related to the choice of similarity measure. The use of other measures may permit less computationally intensive methods to be used for data organization. Topic specification documents are shown to be lexically distant from clusters of relevant documents. The significance of this finding is unclear.

Acknowledgements—It is a pleasure to acknowledge the many helpful comments of Dr. Donna Harman of the National Institute of Standards and Technology and two anonymous reviewers. We also wish to thank Marc St. Gil, Unix Support Director, Academic Computing Services of the University of North Texas for his cooperation in securing the machine time necessary to complete this study. This paper was supported in part by UNT RIG #35396.

REFERENCES

Davidson, M.L. (1983). *Multidimensional scaling*. New York: John Wiley.

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science 45*(1), 12–19.

Dubes, R. & Jain, A. (1979). Validity studies in clustering methodologies. *Pattern Recognition 11*, 235–254.

Goodrum, A. (1997) Evaluation of Text-Based and Image-Based Representations for Moving Image Documents. http://people.unt.edu/ ~ agoodrum/research.

Goodrum, A. (1997). Evaluation of text-based representations for moving image documents. Unpublished Ph.D. Dissertation, University of North Texas School of Library and Information Sciences, Denton, TX.

Gupta, A. & Jain, R. (1977). Visual information retrieval. *Communications of the ACM 40*(5), 71–79.

Harman, D. (1993). Data preparation. In R. H. Merchant (Ed.), *Proceedings of the TIPSTER Text Program—Phase I* (pp. 17–31). San Francisco, CA: Morgan Kaufman.

Harman, D. (1994). Overview of the second text retrieval conference, (TREC-2). In D.K. Harman (Ed.) *The Second Text Retrieval Conference (TREC-2)* (pp. 1–20) Gaithersburg, MD: National Institute of Standards and Technology.

Harman, D. (1995). Overview of the third text retrieval conference (TREC-3). In D.K. Harman (Ed.), *The Third Text Retrieval Conference (TREC-3* (pp. 1–19) Gaithersburg, MD: National Institute of Standards and Technology.

Harman, D. (1996). Overview of the fourth text retrieval conference (TREC-4). In D.K. Harman (Ed.) *The Fourth Text Retrieval Conference (TREC-4)* (pp. 1–23) Gaithersburg, MD: National Institute of Standards and Technology.

Hearst, M. & Pedersen, J. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In H. Frei *et al.* (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 76–84). Konstanz, Germany: Hartung-Gorre.

Hendry, D. (1995). *Dynamic econometrics*. Oxford: Oxford University Press.

Katter, R. (1967). *Study of document representations: multidimensional scaling of indexing terms*. Santa Monica, CA: System Development Corporation.

Katter, R., Holmes, E. & Weis, R. (1971). *Interpretive overlap among document surrogates: effects of judgmental point of view and consensus factors*. Santa Monica, CA: System Development Corporation.

Korfhage, R. (1991). To see, or not to see. Is that the query? *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 134–141). Chicago: ACM.

Korfhage, R., Lin, X. & Dubin, D. (1995). VIRI: Visual information retrieval interfaces. In E. Fox, *et al.* (Eds.) *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 3787). Danvers, MA: ACM Press.

Larson, R. (1996). Bibliometrics of the world wide web: an exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the 59th Annual Meeting of the American Society for Information Science* (pp. 71–78). Medford, NJ: Learned Information.

McCain, K. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science 41*(6), 433–443.

Ramsay, R. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika 42*(2), 241–266.

Rorvig, M. (1988) Psychometric measurement and information retrieval. In M.E. Williams (Ed.) *Annual review of information science and technology (ARIST)* (Vol. 23, pp. 157–189).

Rorvig, M. (1998). Scaled structure in visualized TREC data and query feedback. *Information Processing and Management. 34*(2/3), 151–160.

Rorvig, M. & Hemmje, M. (1998). Foundations of advanced information visualization for visual information retrieval systems. *Journal of the American Society for Information Science*, in press.

Rorvig, M., Fitzpatrick, S., Ladoulis, T. & Vitthal, S. (1993). A new machine classification method applied to human peripheral blood leukocytes. *Information Processing and Management 29*(6), 765–774.

Salton, G. & McGill, M. (1981). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Satoh, K., Akamine, A. & Okumura, A. (1996). Improvements on query term expansion and ranking formula. In D.K. Harman (Ed.) *The Fourth Text Retrieval Conference (TREC-4)* (pp. 475–482). Gaithersburg, MD: National Institute of Standards and Technology.

SAS Institute Inc. (1996). *SAS/STAT Software: changes and enhancements through Release 6.11*. Cary, NC: SAS Institute Inc.

Shaw, W., Burgin, R. & Howell, P. (1997). Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing and Management 33*(1), 1–14.

Sparck Jones, K. & Van Rijsbergen, C. (1975). Report on the Need for and the Provision of and 'Ideal' Information Retrieval Test Collection. *British Library Research and Development Report 5266*. Cambridge: Computer Laboratory.

Small, H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science 24*(4), 265–269.

Text Navigator: http://direct.boulder.ibm.com/dss/cas/cas.htm and http://9.101.131.144/ecamdemo.htm.

ThemeMedia: http://www.smaby.com/thememedia.html.

Van Rijsbergen, C. (1989). Towards an information logic. *Research Report CSC/89/R8*. University of Glasgow: Dept. of Computing Science.

Veerasamy, A. & Belkin, N. (1996). Evaluation of a tool for visualization of information retrieval results. In H. Frei *et al.* (Ed.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 85–92). Konstanz, Germany: Hartung-Gorre.

Voorhees, E. (1985). The cluster hypothesis revisited. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 188–196).

Weis, R. & Katter, R. (1967). Multidimensional Scaling of Documents and Surrogates. *Technical Memorandum SP-2713.* Santa Monica, CA: System Development Corporation.

White, H. & Griffith, B. (1981). Author cocitation: A literature measure of intelligent structure. *Journal of the American Society for Information Science 32,* 163–171.