Images of Similarity: A Visual Exploration of Optimal Similarity Metrics and Scaling Properties of TREC Topic-Document Sets

Mark Rorvig
School of Library and Information Sciences, The University of North Texas, P.O. Box 311068, Denton, TX 76203-1068. mrorvig@jove.acs.unt.edu; http://archive.lis.unt.edu:2025/resume

Multiple similarity measures for five TREC topic-document sets from the LDC TREC Collection Disk 1 are derived from the full text of documents. Each measure on each set is scaled using SAS MDS under ordinal, interval, and MLE assumptions. The resulting 75 permutations are ploted. It is suggested that cosine-vector and overlap measures for similarity appear to recover optimal data relationships among the documents of the five sets. MLE assumptions appear to be required to model the data adequately.

Introduction

A few years ago, Bob Korfhage (Korfhage, 1995) threw down a challenge to IR researchers to explore the differences among various similarity measures in the calculation of relative distance measures among document sets. He noted that the subspace consisting of documents "close" to a given document is very different for different kinds of measures, even though the documents in question do not themselves change. Korfhage had a strong incentive to explore this issue because understanding these differences is critical to understanding the different performance issues arising among Visual Information Retrieval Interfaces (VIRIs) (Korfhage, Lin & Dubin, 1995); a rapidly developing field in which he has played a seminal role.

There are now quite a few VIRI systems of which Gupta and Jain (1997), Hemmje, Kunkel, & Willett (1994), Kohenen (1989), Lin (1997), Olsen, Korfhage, Sochats, Spring, & Williams (1996), Rorvig and Wilcox (1997), and Wise et al. (1995) are indicative. In all these approaches, document distances scaled by one measure or another are critical aspects of user navigation within a document space.

The notion of a "document space" itself is quite old in IR research, dating back to early work by Salton (1971) and also by Koll (1979). However, the present concept is more directly enunciated by the work of Crouch (1986), in which the use of scaling techniques to display the inner content of similarity matrices derived from document sets was demonstrated for the first time.

In 1996, at the ACM SIGIR Annual Meeting in Zurich, it was suggested at a visualization workshop that the TREC IR Test Collection (Harman, 1993, 1994, 1995, 1996) should be used to evaluate VIRI issues (Rorvig & Hemmje, 1998). As a result of this decision, some work has been done that demonstrates the promise of this approach. In Rorvig & Fitzpatrick (1997), for example, highly interrelated document areas of the visual field were found to be closely associated with the presence of relevant documents. Moreover, relevant documents drawn from these dense areas when used in document surrogate query feedback experiments were shown to significantly increase the number of retrieved relevant documents over relevant document surrogate queries drawn from the less dense areas of the visual field (Rorvig, 1997).

The TREC topic-document subsets themselves, as observed and judged by analysts, are unique in several respects. However, the most important of these in the context of this study is the effect on document similarity created by the item evaluation process (Harman, 1993, 1996) intrinsic to the test collection assembly. Documents available from the source text collections of TREC (e.g., Associated Press newsfeeds, Wall Street Journal articles, etc.) were far too numerous to be judged with respect to a topic specification. Thus, a "pooling" method was used to narrow the number of documents that an analyst would need to consider for relevance to a topic.

Originally suggested as a technique for test collection creation by Sparck Jones and Van Rijsbergen (1975), the procedure was operationalized by engaging the services of TREC/Tipster participants to load the entire set of documents in their systems and then use the topic specification as a query. Each participant, about 20 in all, then provided the

Received January 8, 1998; revised June 4, 1998; accepted June 4, 1998.

TREC project with a list of the top 100 items ranked by each system with respect to the query. From these lists, the refined list of review documents for the analysts were compiled.

The effect of this procedure was to create a group of highly interrelated documents with respect to each topic. The significance of this effect for this study is that a normative view of the document space, by Crouch's definition, becomes possible. No matter what exact criteria were used by the judges to determine relevance, and no matter what ranking algorithms were used by the participant systems, two assumptions of high construct validity emerge: first, relevant documents should be more alike one another than nonrelevant documents in their token composition; second, nonrelevant documents should be less like one another and less like relevant documents in their token composition.

The expected document space created under these assumptions by visualizing these topic-document sets thus emerges as a "bulls-eye" composition in which the relevant documents should be centered in the visual field and the nonrelevant documents dispersed around this center in the rough shape of a penumbra. This study is thus a search for the similarity measures and ordering assumptions that most faithfully recover the conditions under which the data were selected from their surrounding collections and judged for relevance by human analysts.

Understanding which similarity measures and ordering assumptions are most likely to produce this normative composition is critical to VIRI testing. The presumption among all VIRIs is that human attention is conserved by presentation to a user of a set of documents in a document space that visually encompass the document interrelationships. The user should be able to see at a glance the most highly related documents, as opposed to the cognitively complex alternative of perusing a long list of items in order to draw the same conclusions. If, however, document relationships present in the collection are not modeled by the statistical methods of a particular VIRI, the gains in human attention and cognition will be attenuated, and the purpose of VIRI technology mitigated.

Further, as discussed in more detail later in this paper, the choice of measures and ordering assumptions pose strong conditions on the degree and nature of the computational power needed to create visual fields of document relationships. Indeed, it is one of the more dismal conclusions of this study that significant computational power may be required for such visual relationships to be presented to users in an interactive IR environment.

Procedures

TREC data were obtained for this study by purchase of Volume I of the TREC six-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 the ZIFF collection subset (such correspondences exist on succeeding TREC volumes.) 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.

Five topic-document sets were culled from the Volume I collection, specifically, topic-document sets 1, 3, 5, 7, and 9. These data were then placed in separate directories on a Sun Sparcstation 4. The number of documents in each set ranged from a low of 421 for document set 9 to a high of 586 for document set 1. A program was then written in C++ to produce five similarity matrices for each set. Because this study was exploratory in nature, no attempt to employ stoplists, stemmers, or term weights was made. (Some effects of these techniques on TREC visual fields is reported in Rorvig, Sullivan, and Oyarce (1998).

The similarity measures were chosen from those previously identified as most common in IR research by Salton and McGill (1983, pp. 200–204) and Meadows (1992, pp. 201–204). These five measures are identified respectively as Dice, Jaccard, cosine, overlap, and asymmetric measures and identified as SIM_1 though SIM_5, respectively, throughout this paper. Of these measures, the experimental literature has most frequently dealt with Dice (Griffiths, Luckhurst, & Willett, 1986), Jaccard (Van Rijsbergen, 1989; Kopcsa & Schiebel, 1998), and cosine measures broadly applied by Salton. Formulas for each of these measures appear below.

$$\mathrm{SIM}_1(\mathrm{DOC}_i,\mathrm{DOC}_j) = \frac{2\left[\sum_{k=1}^{t} \mathrm{TERM}_{ik} \cdot \mathrm{TERM}_{jk}\right]}{\sum_{k=1}^{t} \mathrm{TERM}_{ik} + \sum_{k=1}^{t} \mathrm{TERM}_{jk}} \tag{1}$$

$$\mathrm{SIM}_2(\mathrm{DOC}_i,\mathrm{DOC}_j) = \frac{\sum_{k=1}^{t} \mathrm{TERM}_{ik} \cdot \mathrm{TERM}_{jk}}{\sum_{k=1}^{t} \mathrm{TERM}_{ik} + \sum_{k=1}^{t} \mathrm{TERM}_{jk} - \sum_{k=1}^{t} (\mathrm{TERM}_{ik} \cdot \mathrm{TERM}_{jk})} \tag{2}$$

$$\mathrm{SIM}_3(\mathrm{DOC}_i,\mathrm{DOC}_j) = \frac{\sum_{k=1}^{t} (\mathrm{TERM}_{ik} \cdot \mathrm{TERM}_{jk})}{\left[\sum_{k=1}^{t} (\mathrm{TERM}_{ik})^2 \cdot \sum_{k=1}^{t} (\mathrm{TERM}_{jk})^2\right]^{1/2}} \tag{3}$$

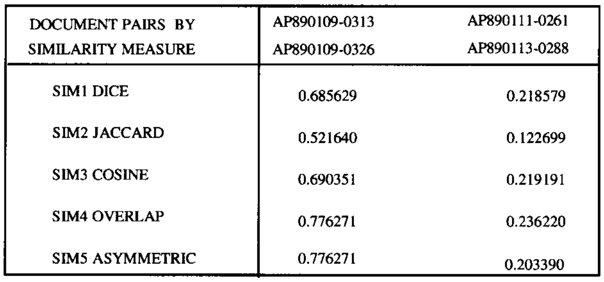
$$\mathrm{SIM}_4(\mathrm{DOC}_i,\mathrm{DOC}_j) = \frac{\sum_{k=1}^{t} (\mathrm{TERM}_{ik} \cdot \mathrm{TERM}_{jk})}{\min\left(\sum_{k=1}^{t} (\mathrm{TERM}_{ik})^2,\ (\mathrm{TERM}_{jk})^2 \right)} \quad (4)$$

$$\mathrm{SIM}_5(\mathrm{DOC}_i,\mathrm{DOC}_j) = \frac{\min \sum_{k=1}^{t} (\mathrm{TERM}_{ik},\ \mathrm{TERM}_{jk})}{\sum_{k=1}^{t} \mathrm{TERM}_{ik}} \quad (5)$$

All five of these similarity measures produce different measures of document closeness, as measured by the degree to which two documents i and j share text tokens k. All produce higher values of document closeness as the number of tokens in any pair of documents increases. The most significant difference among them, however, is the method by which the denominator of each formula treats differences in document length. In SIM_1 (Dice), for example, document length is simply additive; in SIM_2 (Jaccard) the document length is additive but reduced by the total number of terms two documents have in common; in SIM_3 (cosine), document length is magnified by squaring terms that occur multiple times in a document vector, summing the vector products and taking their square root; in SIM_4 (overlap) common tokens between two documents are reduced to the token count for corresponding term vectors containing the least occurrences of that term; while in SIM_5 (asymmetric), only the number of terms in document i of each pair of documents is considered.

These calculations result in different values for the same pairs of documents as shown in Figure 1 below. The differences are slight between the first two documents because they deal essentially with the same topic (a potential hostile takeover of the same company), while for the second pair, lower values result because the articles have little in common. The four documents differ little in length. Scores for the overlap and asymmetric measures are identical for the first two documents because of their high degree of token similarity. (Documents are reproduced in Appendix 2.)

Each similarity matrix produced by each method was then scaled using multidimensional scaling (MDS) as implemented in SAS (SAS Institute, 1996) installed on a Sun Ultra Enterprise 5000 class machine running at 167 MHz with 4 CPUs and 1 Gb of RAM. MDS was used for this analysis due both to the suggestions of Crouch (1986), who first employed the technique and to its wide use and rich history throughout IR research [for examples see: Katter (1967), Weis & Katter (1967), Katter, Holmes, & Weiss (1971), Small (1973), White & Griffith (1981), Rorvig (1988), McCain (1990), Rorvig, Fitzpatrick, Ladoulis, & Vitthal (1993), Larsen (1996), Goodrum (1997), Kopcsa & Schiebel (1998). Background material and comprehensive bibliographies for MDS may be found in Rorvig (1988). Goodrum (1997) updates the literature in the area during the period between the two publications.



DOCUMENT PAIRS BY SIMILARITY MEASURE	AP890109-0313 AP890109-0326	AP890111-0261 AP890113-0288
SIM1 DICE	0.685629	0.218579
SIM2 JACCARD	0.521640	0.122699
SIM3 COSINE	0.690351	0.219191
SIM4 OVERLAP	0.776271	0.236220
SIM5 ASYMMETRIC	0.776271	0.203390

FIG. 1. Table of similarity values produced by five measures on two pairs of documents from TREC topic-document set 1.

Each matrix was scaled in MDS at three levels of assumptions about the implicit a priori organization of the data regardless of its similarity measure: ordinal, interval, and, if interval, than with error terms distributed lognormally. Formally, the strictest case of the classic MDS model may be expressed as

$$\delta_{i,j} = d_{i,j} = \left[\sum (x_{i,r}. \, . \, .x_{j,r})^2 \right]^{1/2} \quad (6)$$

where $\delta_{i,j}$ is the similarity (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. Any set of data points may be solved perfectly by this equation, provided that the number of dimensions is equal (that is, $r = N - 1$) to the number of objects measured. The goal of MDS, however, is to reduce the dimensionality of the data to the smallest possible number, while preserving the initial order of the direct measurements. Hence, the stronger the assumptions about the underlying measurements, the more difficult the dimensional reduction becomes and the greater the error term in the fitting of the original measurements to proximities. It makes little sense to construe similarity measures between two documents as nominal measurements (e.g., similar or not similar), but it does make sense to test the proposition that such measures are ordinal in level.

When document similarity measures are considered, one can be sure that large differences in similarity values between two pairs of documents are meaningful representations of token composition differences. However, to proceed to the next inferential step and assume that such differences represent real numbers on a continuum with a common unit of measure is a strong one. A pair of documents may be referred to as more or less similar than another pair, but such differences may not represent real values. The consequences for scaling data at an assumed ordinal level are that the distances recovered from the similarity values may be fitted so as to produce an error free pattern among all pairwise measures in a dataset. The ordinal solution in MDS is also the most efficient of the various computational implementations because the least restrictions are placed on the algorithm to recover the initial configuration of data.

The stronger assumption regarding document similarity measures is that such measures are interval or ratio in nature, and represent fixed points on a continuous scale. Such assumptions thus also introduce the problem of distributions of the observations about a mean and the consequent error term arising from variability in the measurements. If little variability is present among the distributions of values of similarity, that is, their distribution is approximately normal, MDS performs a transformation from similarity measures to distances with little difficulty. However, within the TREC datasets, as observed in Rorvig and Fitzpatrick (1997), document types (e.g., Federal Register documents) tend by virtue of their length and compositional style to be much more like one another than like other document types such as news feeds and the like. Thus, matrices of similarity measurements of TREC datasets tend to have quite variable distributions for which the arithmetic mean used by MDS in distance transformations is unreliable.

This type of condition in data is not unique to document measurements, however, and in the late 1970s Ramsay (1977) proposed a maximum likelihood method (MLE) to calculate a mean of observed data by assuming a lognormal distribution of values because the lognormal distribution has a constant ratio of the standard deviation to the mean. 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

$$P(\delta|d,\, \sigma^2) = (2\pi)^{1/2}(\sigma\delta)^{-1}\exp\left[\frac{-\ln(\delta|d)}{2\sigma^2}\right] \qquad (7)$$

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

$$\ln L = \sum \sum \ln P(\delta_{i,j}|d_{i,j},\sigma^2). \qquad (8)$$

This procedure was implemented as a standard option in SAS MDS in 1986 (Ramsay, 1986). Nevertheless, MLE is an iterative numerical method carrying a high computational overhead. Dataset 1, for example can be transformed by MDS from raw measurements to proximities under ordinal assumptions on the Sun Explorer 5000 in about 10 CPU minutes, while the MLE implementation requires from 1.2 to 2 hours, depending on the size of the initial input matrix.

These treatments thus result in 75 plot representations of the five datasets appearing as Figures 2–6 in Appendix 1 of this paper.

Results

Among the 75 treatments, the most striking feature is the great deal of variability present in the visual images of these data. On each page of Figures 2–6, only one dataset is present, yet 15 different images emerge. Some general patterns are:

1. Regardless of the choice of similarity measure and with only two possible exceptions (Dice and Jaccard measures in ordinal to interval level treatments in datasets 5 and 7), tighter groupings emerge as more variability is assumed to be present in the data and MDS treatments move from ordinal to MLE requirement assumptions. (Moreover, the scale of the data also changes from treatment to treatment, but without revealing any specific patterns of variation.)
2. The Dice and Jaccard representations produce the most similar images between ordinal and interval treatments. In some instances, the visual plots (accentuated by the size of frame reproduction for publication) appear to be identical, especially for dataset 7 in Figure 5. However, a UNIX diff command applied to the plot files reveal that no plot points are common among the images, and the visual similarity does not extend to the internally recovered observations.
3. In all cases, the asymmetric measure treated under interval level assumptions by MDS results in an arching pattern, due to the single document denominator used in this measure for all document pairs. Under ordinal level treatments, this effect on asymmetric measures is reduced by the MDS error free solution, but not eliminated.
4. The overlap and cosine similarity measures produce the most regular tightening of pattern dispersion among all five datasets as treatments levels go from ordinal to MLE assumptions by MDS.
5. In half of all similarity measures treated by MDS under ordinal and interval level data assumptions, islands of data appear. Analysis of these separated, tight, nonrelevant groups indicates that they are composed entirely of Federal Register documents. This effect is most pronounced in dataset 1 of Figure 2 and dataset 9 of Figure 6. Federal Register documents within the TREC collection tend to be much longer than other document types, indicating that document length poses a problem in lower level scaling treatments.
6. Only the cosine and overlap measures under MLE level assumptions reproduce the “bullseye” pattern expected in the data. The cosine measure under MLE level treatment produces the tightest grouping of data with the least dispersion of relevant documents. However, the overlap measure more consistently creates an ovoid pattern. Differences between these two measures may merely amount to a matter of aesthetics, or more importantly, may indicate a valuable measure of term importance incorporated in the cosine measure.
7. The regularities of pattern noted in (6) on relevant documents indicates high orderliness and consistency in the assigned binary relevance judgments by human analysts. This striking degree of regularity is unexpected, given the reservations concerning relevance judgment consistency expressed by Cooper, Gey, and Chen (1993), Harter (1996), and Rorvig (1988, 1990).

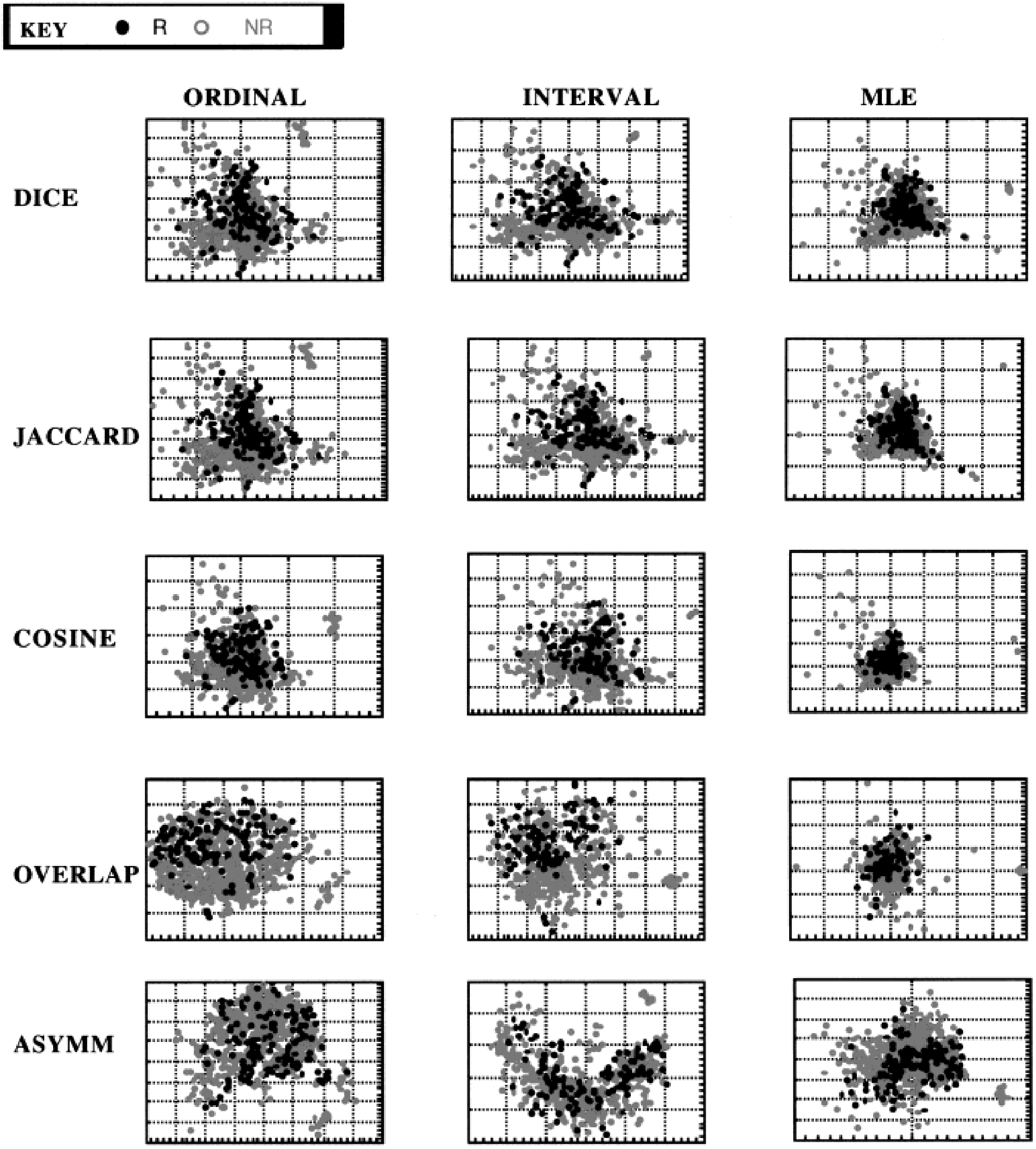


FIG. 2. TREC topic-document dataset 1 by five similarity measures and three scaling methods ($n = 585$).

Discussion

Confronted by these data, one is tempted to repeat the homely (and evasive) words of Pontius Pilate at the inquisition of the Christ and say, “What is Truth?” That the same data produce so many different visual patterns under different measures and ordering treatments is both disturbing and reassuring at the same time.

On the reassuring side, it is clear from this study that visualization as an analytical tool, irrespective of its use in VIRIs, demonstrates great power in unraveling cause and effect in retrieval. It is not surprising, for example, given the cosine measure’s long history of performance in SMART and SIRE implementations, that it should produce such outstanding results in recovering the appropriate visual shape from these TREC topic-document subsets. But it is very surprising that the overlap measure should perform as well as the cosine measure. Indeed, the overlap measure may be a much undervalued approach to query-document ranking in retrieval, and clearly, more efforts should be made to exploit it.

It is worth commentary that the cosine measure is more sensitive than the overlap measure to the presence of multiple word occurrences in texts due the way term vectors are processed in the denominator of the equation. An IDF weight might alter the shape of the cosine distribution into the oval one of the overlap. If this were the case, then the overlap measure would actually be preferred, because it would be more computationally efficient.

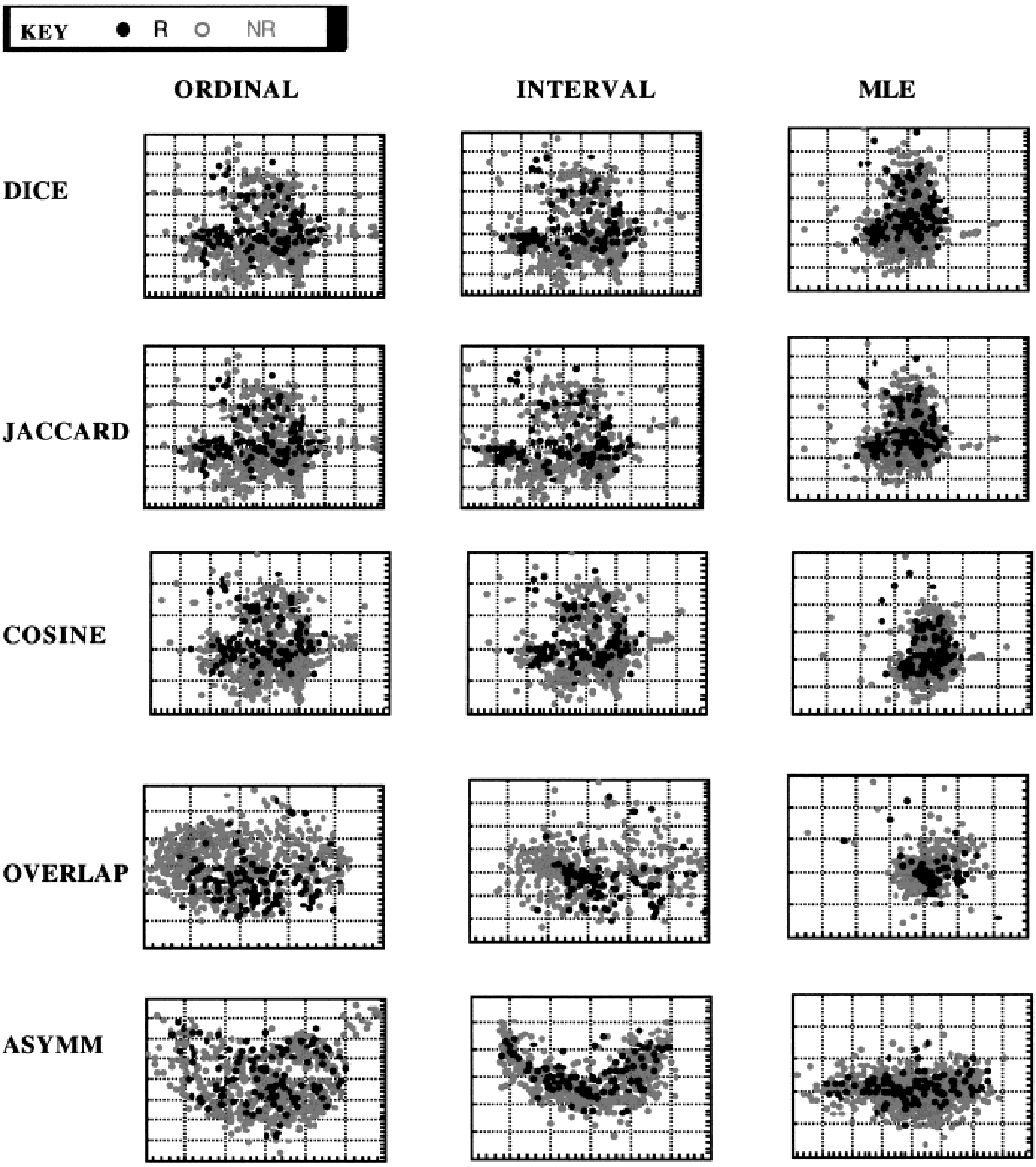


FIG. 3. TREC topic-document dataset 3 by five similarity measures and three scaling methods ($n = 531$).

Additionally, the performance of both the overlap and cosine measures in ordering these documents and folding long ones into the central core suggests a valuable implementation strategy for VIRIs to incorporate the findings of Singhal, Salton, Mitra, and Buckley (1996), who notes that document length is not independent of relevance, and that longer documents tend to be more relevant than shorter ones. In a VIRI display, the longer documents should not be separated into islands (as they frequently are in other measures).

On the disturbing side, and especially germane to VIRI technology, the conclusion that an MLE approach to error distribution in these data is a prerequisite to their proper organization for retrieval display is disheartening because it suggests that so much computational power may be required to implement an interactive VIRI that the technology may be years away from broad scale use and availability. There may be some hope that these computational processes may be optimized, however, by use of relatively new techniques for document representation, such as the document feature vector approaches taken by Lewis (1992), Lewis, Schapire, and Callan (1996), Apte, Damerau, and Weiss (1994), and Hearst, Pedersen, Pirolli, Schutze, Grefenstette, and Hull (1994). Evidence in favor of this approach is provided by Kopcsa and Schiebel (1998), who are able to show that although an iterative approach is still required when term

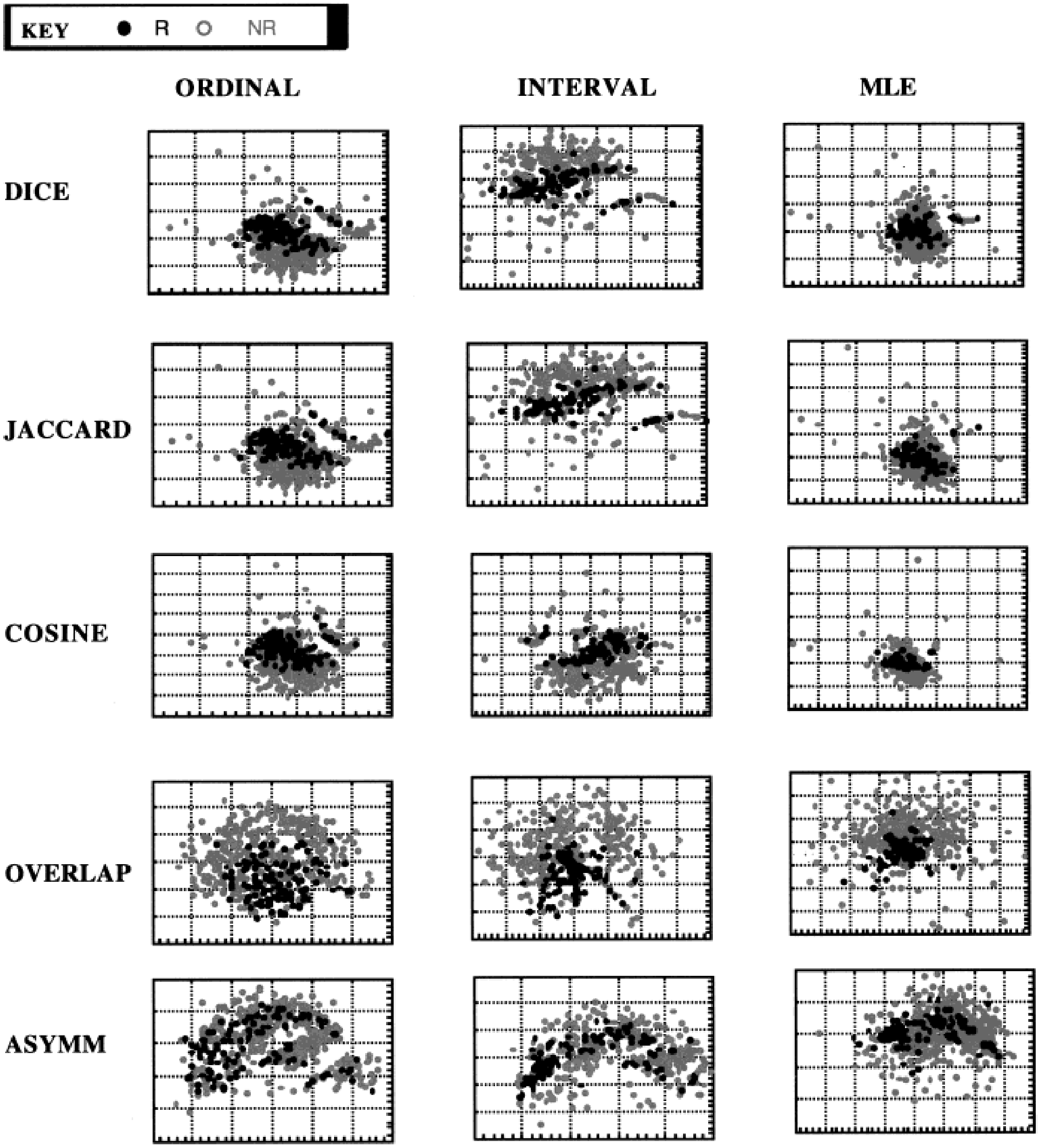


FIG. 4. TREC topic-document dataset 5 by five similarity measures and three scaling methods ($n = 464$).

distributions are not normally distributed, nearly as good a recovery of the initial data fit is possible without resorting to MDS. Kopcsa and Schiebel (1998), however, choose the Jaccard measure, which does not perform well in this study, possibly indicating that the superiority of their results over MDS may simply be an artifact of their choice of initial similarity measures.

Finally, these visualizations suggest that the variability of judgments of relevance may be less a function of judgment processes and more attributable to a failure to appropriately organize and scale data (although much room remains for the attribution of judgment variability to individual differences in judges as suggested in Harter (1996) and Saracevic (1991). Nevertheless, order is a precious commodity wherever and however it is found, because the discovery of such order frequently leads to greater control over experimental processes.

Conclusions

For IR collections of heterogeneous length and token composition, that is, nearly all collections, visual presentations within a VIRI context of static or interactively retrieved documents cannot be presented under assumptions of ordinal measurements if MDS or related techniques are used for visualization. This finding will remain true, regardless of the initial similarity measure chosen for similarity calculation. Moreover, if MDS is used as the scaling method for visual presentation, interval level assumptions about the

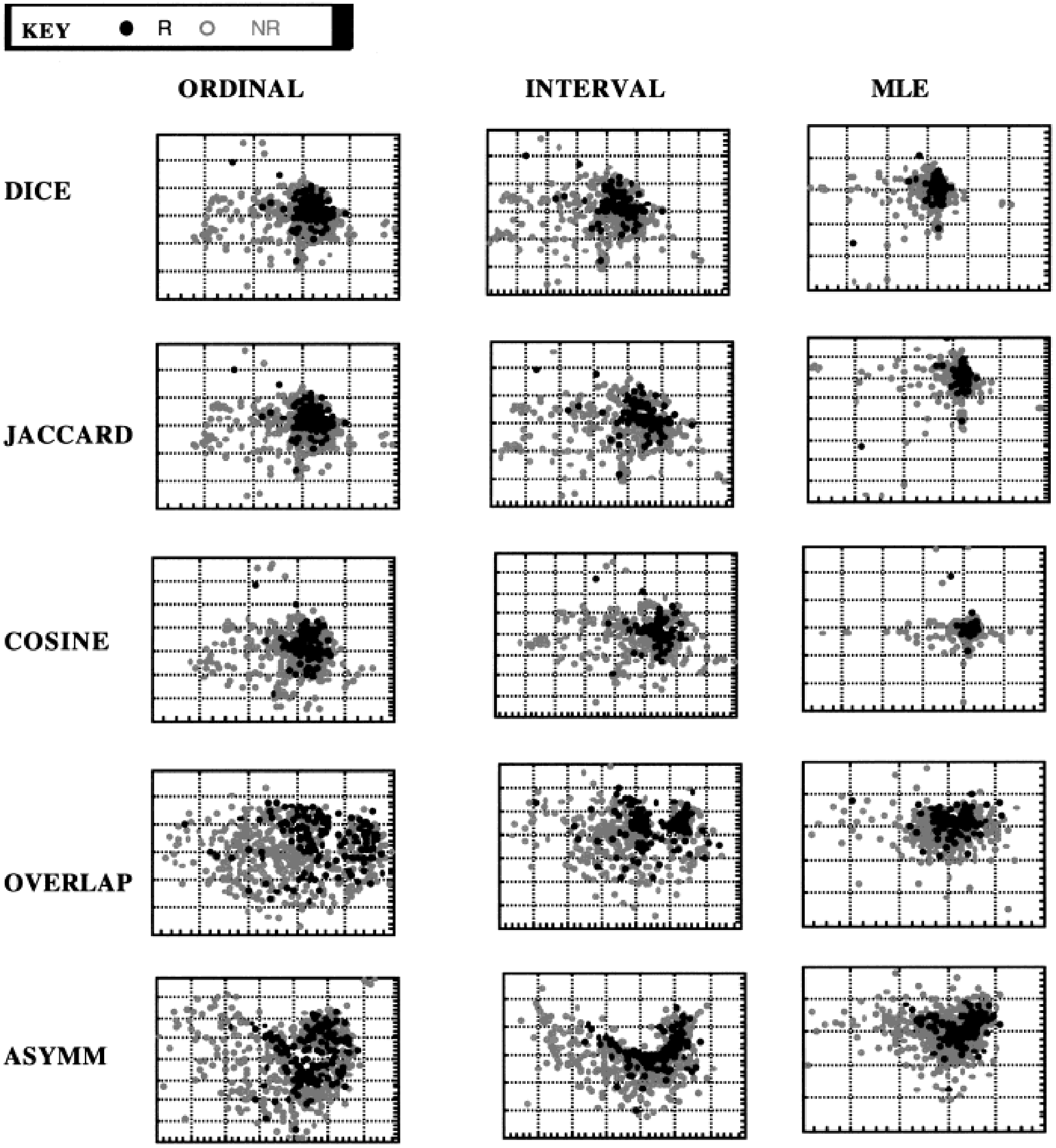


FIG. 5. TREC topic-document dataset 7 by five similarity measures and three scaling methods ($n = 491$).

initial similarity measures should also be processed, assuming a lognormal distribution of error terms in the fitting of document similarity to document distances.

Of the various similarity measures considered in this paper, only cosine and overlap (SIM_3 and SIM_4) measures satisfactorily recover known data characteristics in the TREC IR test collection. The overlap measure performs very well in the context of this study and deserves further attention as a general approach to query-document partial matching in conventional IR systems.

The regularity of visual ordering of relevant documents in this study for the ideal treatments of cosine or overlap measures and MLE MDS proximity calculation is unexpected. Broader analysis of more TREC topic-document data sets is encouraged to determine whether this organizing effect arises as strongly over a wider population of TREC topic-document sets.

Acknowledgments

This research was supported in part by UNT University Research Institute Grant #60207 and UNT RIG #35396. It is a pleasure to acknowledge the contribution of Dr. Donna Harman of the National Institute of Standards and Technology, whose discussion with the author in the Summer of 1997 on the pooling technique used in the TREC Test Collection construction process led to this study. Appreciation is also due to Mr. David Zych, a senior computer science major at the University of North Texas, who contributed the similarity matrix C++ code used in this study.

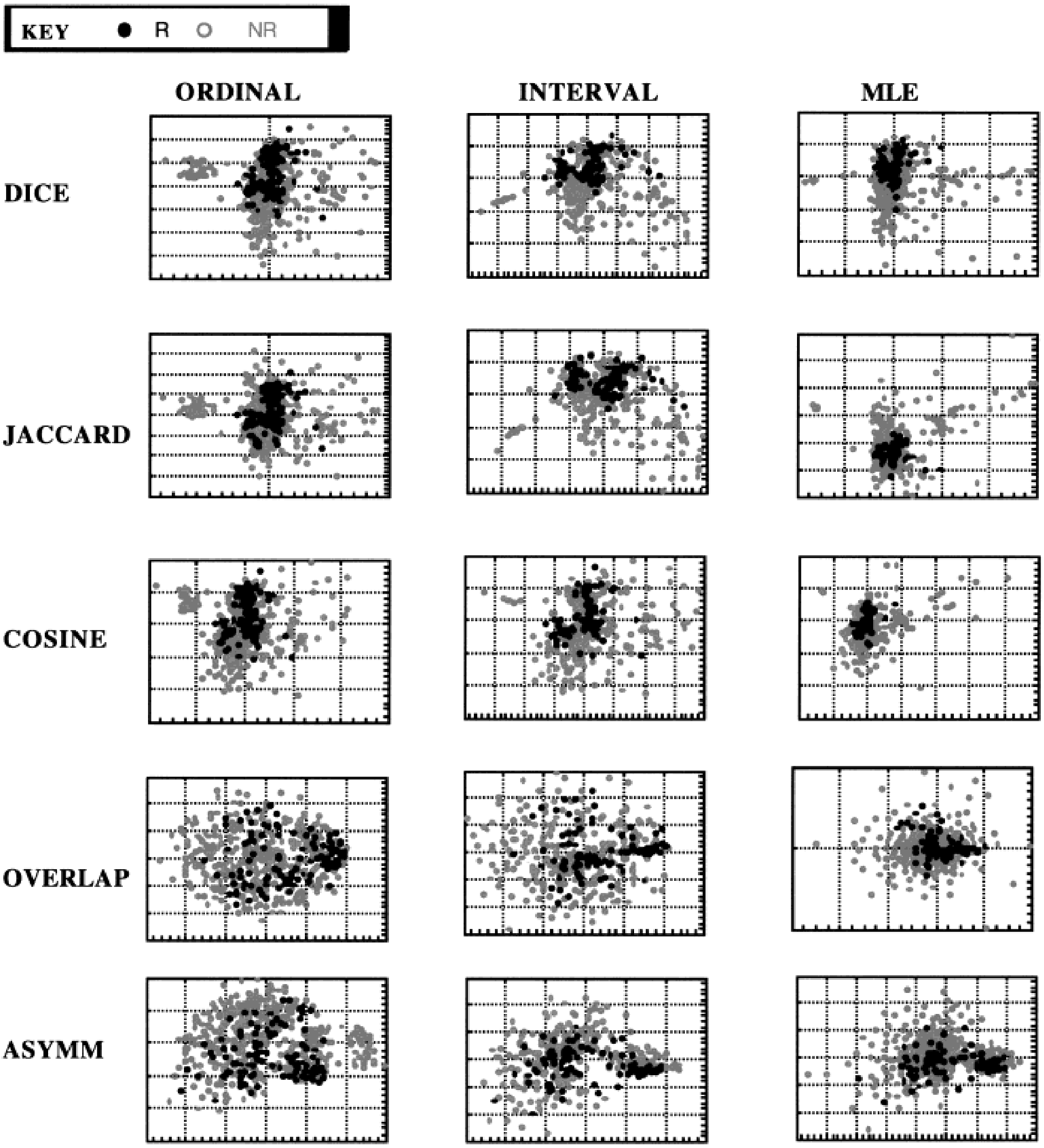


FIG. 6. TREC topic-document dataset 9 by five similarity measures and three scaling methods ($n = 421$).

Finally, the author gratefully acknowledges the contributions of two anonymous reviewers to both the clarity and accuracy of this study.

References

Apte, C., Damerau, F., & Weiss, S. (1994). Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), 233–251.

Cooper, W.S., Gey, F.C., & Chen, A. (1993). Probabilistic retrieval in the TIPSTER collections: An application of staged logistic regression. In D.K. Harmon (Ed.), The First Text Retrieval Conference (TREC-1) (p. 80). Gaithersburg, Maryland: Dept. of Commerce, NIST Special Publication 500-207.

Crouch, D.B. (1986). The visual display of information in an information retrieval environment. In F. Rabitti (Ed.), Proceedings of the ninth annual international ACM SIGIR conference on research and development in information retrieval (pp. 58–67). Pisa, Italy: ACM.

Dubin, D.S. (1997). Structure in document browsing spaces. Unpublished Ph.D. Dissertation, Department of Information Science, University of Pittsburgh, Pittsburgh, PA.

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

Griffiths, A., Luckhurst, H., & Willett, P. (1986). Using interdocument similarity information in document retrieval systems. Journal of the American Society for Information Science, 37(1), 3–11.

Gupta, A., & Jain, R. (1997). 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—PhaseI (pp. 17–31). San Francisco: 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.

Harter, S.P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47(1), 27–49.

Hearst, M., Pedersen, J., Pirolli, P., Schutze, H., Grefenstette, G., & Hull, D. (1994). Xerox TREC4 site report. In Proceedings of the fourth text retrieval conference.

Hemmje, M., Kunkel, C., & Willett, A. (1994). LyberWorld—A visualization user interface supporting full text retrieval. In Proceedings of the seventeenth annual international ACM SIGIR conference on research and development in information retrieval (pp. 249–258). Dublin, Ireland: ACM.

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.

Koll, M.B. (1979). The concept space in information retrieval systems as a model of human concept relations. Unpublished Ph.D. Dissertation, School of Information Studies, Syracuse University, Syracuse, NY.

Kohonen, T. (1989). Self-organization and associative memory (3rd ed.). New York: Springer Verlag.

Kopcsa, A., & Schiebel, E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. Journal of the American Society for Information Science, 49(1), 7–17.

Korfhage, R. (1995). Some thoughts on similarity measures. SIGIR Forum, 29, 8.

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. 377). Danvers, MA: ACM.

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.

Lewis, D. (1992). Representation and learning in information retrieval. PhD. Thesis, Department of Computer and Information Science, University of Massachusetts at Amherst.

Lewis, D., Schapire, R., & Callan, J. (1996). Training algorithms for linear text classifiers. In 19th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Lin, X. (1997). Map displays for information retrieval. Journal of the American Society for Information Science, 48(1), 40–54.

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

Meadows, C.T. (1992). Text information retrieval systems. San Diego: Academic Press.

Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., & Williams, J.G. (1996). Visualization of a document collection: The VIBE system. Information Processing and Management, 29(1), 69–81.

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

Ramsay, J.O. (1982). Some statistical approaches to multidimensional scaling data. Journal of the Royal Statistical Society (A), 145, 285–312.

Ramsay, J.O. (1986). The MLESCALE procedure. In SUGI Supplemental Library User's, Version 5. Cary, NC: SAS Institute.

Rorvig, M. (1988). Psychometric measurement and information retrieval. In M.E. Williams (Ed.), Annual Review of Information Science and Technology (ARIST), 23, 157–189.

Rorvig, M. (1997). Scaled and visualized structure in TREC topic-document subsets and query feedback [Accepted for publication in Information Processing and Management, 1997.]

Rorvig, M., & Fitzpatrick, S. (1997). Visualization and scaling of TREC topic-document sets [Accepted for publication in Information Processing and Management, 1997.]

Rorvig, M., & Hemmje, M. (1997). Foundations of advanced information visualization for visual information (retrieval) systems [Forthcoming in ACM SIGIR Forum, Spring, 1998.]

Rorvig, M., & Wilcox, M. (1997). Visual access to special collections. Information Technology and Libraries [Accepted for publication in the September, 1997 issue.]

Rorvig, M., Sullivan, T., & Oyarce, G. (1998). A visualization case study of feature vector and stemmer effects on TREC topic-document subsets [Accepted for publication in the Proceedings of the 1998 annual meeting of the American Society for Information Science, Information Access in the Global Information Economy, October 25–29, 1998, Pittsburgh, PA.]

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. (1971). The SMART retrieval system—Experiments in automatic document retrieval. Englewood Cliffs, NJ: Prentice-Hall.

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

Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. Proceedings of the Annual Meeting of the American Society for Information Science, 28, 82–86.

SAS Institute Inc. (1996). SAS/STAT Software: Changes and enhancements through release 6.11. Cary, NC: SAS Institute Inc.

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. Information Processing and Management, 32(5), 619–633.

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

Sparck Jones, K., & Van Rijsbergen, C. (1975). Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory: University of Cambridge.

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

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.

Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In Proceedings, Information Visualization, October 30–31, 1995, Atlanta, GA, USA. Los Alamitos, CA: IEEE Computer Society Press.

Appendix 2 Documents noted in Figure 1

<DOCNO>AP890109-0313</DOCNO>
<FILEID>AP-NR-01-09-89 1035EST<FILEID>
<FIRST>u f PM-Britain-GEC 01-09 0556</FIRST>
<SECOND>PM-Britain-GEC,0578</SECOND>
<HEAD>Government Looks at Possible Bid for British Electronics Giant</HEAD>

<DATELINE>LONDON (AP)</DATELINE>
<TEXT> The government said today that it was looking at a possible bid for the electronics giant General Electric Co. PLC that an international consortium is expected to launch within days. The takeover, which analysts say could be worth between $11.5 billion and $14.2 billion, would be the largest in Britain. The consortium is expected to include Plessey PLC, another electronics company which is the target of a $3 billion hostile takeover bid from GEC and Siemens AG of West Germany, another electronics company. Although no bid for GEC has been formally launched, the Office of Fair Trading has legal powers to look at a bid “in contemplation.” “We really are looking at the situation to see who the participants are involved before we can take real active steps,” said a spokesman for the office, who asked not to be identified. “There hasn’t actually been a statement of intention.” The Office of Fair Trading usually reviews a bid and then makes a recommendation to the trade secretary on whether he should refer it to the monopolies commission for a full investigation. The bid speculation prompted heavy trading in GEC on London’s Stock Exchange by midday Monday. A GEC spokesman, who wasn’t identified in accordance with British practice, called the developments vague and inconclusive but said that a takeover would be fought. “This appears to be a self-interested attempt by the board of Plessey and its advisers to form a consortium to break up GEC and therefore save Plessey,” he said. The possible bid for GEC began to be taken seriously after the investment firm Lazard Brothers and Co. said over the weekend that it had helped form a company called Metsun Ltd. to devise a proposal “which may or may not” lead to an offer for GEC. Metsun is headed by Sir John Cuckney, chairman of helicopter maker Westland PLC, which was at the center of a 1986 takeover controversy that prompted the resignation of two British Cabinet ministers. Metsun was talking to possible partners both in Britain and abroad, Lazard Brothers said, without identifying them. French electronics company Thomson-CSF said it was considering joining the consortium. Meanwhile, Barclays Bank PLC confirmed it was putting together a $6.2 billion syndicated loan to help finance such a bid. GEC Managing Director Lord Weinstock said in a television interview that his company dropped Barclays Bank as one of its banks because the bank had “behaved in a way that was not quite right.” In addition, Lord Prior, GEC’s chairman, resigned Saturday from the board of Barclays Bank. Potential foreign interest in such an important British manufacturing company is a sensitive political issue, particularly following the controversy surrounding Westland. American helicopter maker Sikorsky Aircraft acquisition of a stake in Westland, Britain’s only helicopter builder, over rival European bidders created the government controversy in 1986. Michael Heseltine, who resigned as defense secretary in opposition to the American bid for Westland, was spearheading demands Monday for Trade Secretary Lord Young to refer the GEC matter to the Monopolies and Mergers Commission “without delay.” The opposition Labor Party has said it will raise the issue in the House of Commons “as a matter of the utmost urgency” when members of Parliament return from the Christmas recess on Tuesday.
</TEXT>
</DOC>
<DOCNO>AP890109-0326</DOCNO>
<FILEID>AP-NR-01-09-89 1441EST</FILEID>
<FIRST>u f AM-Britain-GEC Bjt 01-09 0732</FIRST>
<SECOND>AM-Britain-GEC, Bjt,0761</SECOND>
<HEAD>GEC Says It Would Fight Takeover</HEAD>
<BYLINE>By COTTEN TIMBERLAKE</BYLINE>
<BYLINE>Associated Press Writer</BYLINE>
<DATELINE>LONDON (AP)</DATELINE>
<TEXT> British electronics giant General Electric Co. PLC on Monday accused rival Plessey Co. of mounting “a spoiling tactic” by joining an international consortium that is expected to make a hostile bid for GEC. GEC had joined Siemens AG of West Germany in launching a hostile, $3 billion offer for Plessey in November. Now GEC finds itself the target of a possible counter-bid by a consortium that probably will include Plessey. GEC said it would fight the possible counter-bid and pursue its efforts with Siemens to take over Plessey. Britain’s GEC and U.S.-based General Electric Co. are not related. The government’s Office of Fair Trading said it was studying the possible bid for GEC, which is expected to be launched by a new international consortium within days. That takeover could be worth between $11.5 billion and $14.2 billion, Britain’s largest ever. A counter-bid is known on Wall Street as the Pac Man takeover defense because of its similarity to the eat-or-be-eaten video game. The possibility of GEC being taken over sparked political controversy at the prospect of a large defense-oriented company falling under foreign control. Three years ago, two Cabinet ministers resigned in a controversy over foreign bids for Britain’s only helicopter maker, Westland PLC. GEC said a weekend statement from the investment firm Lazard Brothers &amp; Co. about a possible takeover bid for GEC was “vague and inconclusive.” A GEC spokesman told Press Association, the domestic news agency: “This appears to be a self-interested attempt by the board of Plessey and its advisers to form a consortium to break up GEC and therefore save Plessey, in the short term, regardless of the cost to British industry in terms of lost opportunities and lost jobs.” The spokesman said it was a “spoiling tactic” by Plessey that would be fought. GEC Managing Director Lord Weinstock said in a Sunday television interview that GEC would pursue its bid for Plessey. A GEC spokesman declined to comment to The Associated Press. Plessey also had no comment, a spokesman told the AP. GEC shares jumped 19 pence (34 cents) to finish at 221 pence ($3.91) in heavy trading on London’s Stock Exchange Monday. Plessey shares were down 2 pence (4 cents) at 226 pence ($4) at the close. The possible bid for GEC began to look serious after Lazard Brothers said it had helped form a company called Metsun Ltd. to devise a proposal “which may or may not” lead to an offer for GEC. Metsun is headed by Sir John Cuckney, West-

land's chairman. Metsun was talking to possible international partners, Lazard Brothers said, without identifying them. British news reports said possible participants might include Thomson-CSF of France, American Telephone &amp; Telegraph Co. and General Electric of the United States. Meanwhile, Barclays Bank PLC confirmed it was putting together a $6.2 billion syndicated loan to help finance such a bid. The Office of Fair Trading said it was looking at the possible bid, noting that it has legal powers to review a bid "in contemplation." A spokesman who requested anonymity said the office was examining which participants were involved, adding: "There hasn't actually been a statement of intention." The Office of Fair Trading usually reviews a bid and then makes a recommendation to the trade secretary on whether to refer it to the monopolies commission for full investigation. Rhe acquisition of a stake in Westland by United Technologies Corp.'s Sikorsky division over rival European bidders plunged the government into controversy in 1986. Michael Heseltine, who resigned as defense secretary in opposition to the American bid for Westland, spearheaded demands from politicians and unions for Trade Secretary Lord Young to refer the GEC matter to the Monopolies and Mergers Commission "without delay." After the Office of Fair Trading made its statement, Heseltine said: "I believe the Government is to be congratulated on the speed with which it has acted to give a proper appraisal of the strategic and industrial implications of a breakup of GEC." The opposition Labor Party says it will raise the issue in the House of Commons "as a matter of the utmost urgency" when lawmakers return from the Christmas recess on Tuesday. GEC, which employs more than 100,000 people, had pre-tax profit of $1.2 billion on sales of $9.8 billion in its most recent year.
</TEXT>
</DOC>
<DOCNO>AP890111-0261</DOCNO>
<FILEID>AP-NR-01-11-89 1514EST</FILEID>
<FIRST>u f BC-BiggestDeals 1stLd-Writethru f0132 01-11 0469</FIRST>
<SECOND>BC-Biggest Deals, 1st Ld-Writethru, f0132,0482</SECOND>
<HEAD>Philip Morris-Kraft Deal Tops Fortune's Takeover List For 1988</HEAD>
<HEAD>EDS: SUBS 7th graf to CORRECT to $6.5 billion sted million; picks up 8th graf pvs bg'ng, In third . . . </HEAD>
<DATELINE>NEW YORK (AP)</DATELINE>
<TEXT> Philip Morris Cos.' $12.9 billion acquisition of Kraft Inc. leads a list of the largest corporate takeovers, buyouts and other restructurings completed in 1988, according to a Fortune magazine report. The biggest deal in history_the $25 billion leveraged buyout of RJR Nabisco Inc. by the investment firm Kohlberg Kravis Roberts &amp; Co._was not included in the annual list because the buyout won't be completed until 1989. Also pending at year's end was Grand Metropolitan PLC's $5.7 billion takeover of Pillsbury Co., the magazine said in its report to be published in the Jan. 30 issue. Fortune said the total value of the 50 largest business deals last year was $111.8 billion, easily surpassing the $94.6 billion value of the 50 biggest deals in 1985, the previous record for the magazine's list. Last year's transactions provided at least $687 million in advisory fees to those overseeing the deals_also a record for the list_not counting commissions on related financing, Fortune said, citing public documents and other sources. Despite the publicity over the RJR Nabisco deal, Fortune said most of last year's takeovers and restructurings were done the old-fashioned way: not by corporate raiders wielding borrowed money, but by corporate managers eyeing expansion. The No. 2 deal last year was Canadian developer Robert Campeau's $6.5 billion acquisition of Federated Department Stores Inc., the year's biggest deal involving a foreign company, Fortune said. In third place was the $5.15 billion buyout of Farmers Group Inc., the California insurance company, by British tobacco-retailing conglomerate BAT Industries PLC. The No. 4 deal was the $5.09 billion acquisition of Sterling Drug Inc. by Eastman Kodak Co. In fifth place was Santa Fe Southern Pacific Corp.'s $4.7 billion deal to pay a special dividend to stockholders of the railroad and natural resources company to fend off a takeover bid by the Henley Group. In sixth place was the Kroger Co.'s $3.9 billion deal to pay a special dividend to fight takeover bids for the supermarket chain from Kohlberg Kravis Roberts and the Haft family. The No. 7 deal, the $3.6 billion buyout of papermaker Fort Howard Corp., was the only leveraged buyout in the top 10. In eighth place was the $3 billion acquisition of Triangle Publications Inc., which publishes TV Guide, by Rupert Murdoch's News Corp. The No. 9 deal was the $2.8 billion stock buyback by UAL Inc., the parent company of United Airlines. In 10th place was the largest-ever U.S. acquisition by a Japanese company, Bridgestone Corp.'s $2.6 billion takeover of Firestone Tire & Rubber Co.
</TEXT>
</DOC>
<DOCNO>AP890113-0288</DOCNO>
<FILEID>AP-NR-01-13-89 1925EST</FILEID>
<FIRST>u f AM-Trump-Caesars 01-13 0327</FIRST>
<SECOND>AM-Trump-Caesars,0337</SECOND>
<HEAD>Trump Allowed To Buy More Caesars Stock ATLANTIC CITY, N.J. (AP)</DATELINE>
<TEXT> The Federal Trade Commission has allowed billionaire developer Donald Trump to buy more shares of Caesars World Inc. stock, a Trump official said Friday. Trump has already bought a block of Caesars World stock. In mid-December, he filed a notice under the Hart-Scott-Rodino Act for a determination whether continued purchases would violate antitrust laws. Susan Heilbron, an executive vice president of the Trump Organization, said the FTC notified Trump on Monday that he could buy more stock. "We did not receive any requests for additional information and they raised no questions about the antitrust implications," Ms. Heilbron said. In his request to the FTC, Trump indicated he "has the good-faith intention, subject to

price, availability and necessary approvals," to acquire 15 percent of the stock. He also indicated he may "decide to seek control of (Caesars) through the acquisition of 50 percent (or) more of the outstanding voting securities." Trump, owner of two casino-hotels in Atlantic City and one under construction, has not indicated how much stock he purchased in the rival gaming company. But provisions of the Hart-Scott-Rodino Act require people or companies to file with the trade commission when they amass $15 million worth of stock in another company. The day after Trump officials received the notice from the FTC, Caesars shareholders adopted a shareholder rights program designed to prevent unfriendly takeovers. Under the new plan, if someone buys 20 percent of Caesars stock, or if someone with 10 percent appears only interested in forcing the company to buy back the stock at a premium, or if a person is unqualified to hold a gaming license, then all other Caesars shareholders have the right to buy additional stock at half its market value. Trump has not commented on his involvement with Caesars. Ms. Heilbron also said that even after the FTC ruling, "Mr. Trump would not comment on his plans."
</TEXT>
</DOC>