QUERY EXPANSION SEEN THROUGH RETURN ORDER OF RELEVANT DOCUMENTS

Walter Liggett
National Institute of Standards and Technology
Gaithersburg, MD 20899 USA
walter.liggett@nist.gov

Chris Buckley
SabIR Research, Inc.
Gaithersburg, MD 20878 USA
chrisb@sabir.com

Abstract

There is a reservoir of knowledge in data from the TREC evaluations that analysis of precision and recall leaves untapped. This knowledge leads to better understanding of query expansion as this paper demonstrates. In many TREC tasks, the system response required is an ordered list of 1000 document identifiers. Instead of just using the identifiers to determine the positions of relevant documents in each list, we extract from each list the identifiers of the relevant documents and compare document ordering in these sub-lists. In other words, we consider the return order of relevant documents. We use Spearman's coefficient of rank correlation to compare sub-lists and multidimensional scaling to display the comparisons. Applying this methodology to data from the TREC Query Track, specifically, to system responses to twenty restatements of each of four topics, we show how two systems with query expansion differ from four systems without. We observe return-order variations caused by topic restatement and determine how query expansion affects these variations. For some topics, query expansion reduces the sizes of these variations considerably.

1. INTRODUCTION

Progress in information retrieval (IR) depends on understanding how search results vary with IR system inputs, in particular, the topic (the information need) and the query (the natural language statement that conveys the topic to the IR system). In pursuit of this understanding, TREC evaluations of IR systems elicit system responses (the TREC 1000-document lists) to a variety of topics and for each topic, a variety of queries. TREC data include responses from IR systems with different features and thus, reflect the dependence of feature-related response differences on system inputs. Understanding this dependence can lead to better IR systems. This paper introduces an approach to studying this dependence and applies this approach to comparison of IR systems with and without query expansion.

Computing a performance value from each system response is the customary first step in a TREC analysis. The basis for this is a defining statement of the topic which an assessor then uses to designate some of the documents in the collection as relevant. The document identifiers in the ordered list show the positions in the list occupied by relevant documents. These positions are all that is needed to compute precision and recall measures of performance. In fact, because

there is no designation of degree of relevance for documents in the collection beyond the relevant-irrelevant dichotomy, a performance measure cannot depend on anything except these positions. However, it is not necessary to start with performance values in the analysis of TREC data. Performance values are convenient in that they allow averaging for summarization. Nevertheless, when studying the dependence of system responses on inputs, limiting one's options by insisting on use of a performance-based analysis may hamper the study.

New avenues for analysis are opened when one allows, as the first step, computation of a measure of dissimilarity for each pair of system responses (Banks, et al., 1999). A dissimilarity-based analysis does not preclude a performance-based analysis since the absolute difference between two performance values is a dissimilarity. However, a dissimilarity measure can be computed from the order in which particular documents occur in either the entire list or in the relevant-document part of the list. For example, as detailed in the next section, one can eliminate the irrelevant documents from each list so that one has ordered lists of relevant documents and then compute dissimilarities from these reduced lists by taking into account the actual identifiers in the lists. Such a dissimilarity can be said to depend on the return order of relevant documents. Because such a dissimilarity does not depend on the positions of irrelevant documents in the original list, it is clear that this dissimilarity may reflect aspects of the TREC data that are not portrayed by precision and recall measures.

We apply our dissimilarity-based analysis to the Query Track data of TREC-8 and TREC-9 for the purpose of studying query expansion (Buckley and Walz, 2000). The Query Track data are unique in that they consist of system responses to several restatements of each of 50 topics. As implemented in a well-regarded class of information retrieval systems, the response is computed by first deriving a weighted set of terms (key words) from the original statement of the topic and then matching this query set against the terms in the document collection. The first step may involve a procedure that adds terms to the original query set by examining documents judged particularly relevant in an initial search of the document collection. This procedure is intended to uncover terms pertinent to the need for information that are not in the original statement of this need. Selecting additional terms for the query set may be done by the user in which case the procedure is called relevance feedback (Berry and Browne, 1999) or may be done automatically in which case the procedure is called blind feedback or (automatic) query expansion. Because they include systems with query expansion, the Query Track data allow us to see how systems with and without query expansion handle restatements of information needs.

In thinking about the performance of query expansion, one might suppose that when a system with more effective query expansion is applied to alternative statements of the same need for information, the lists returned would vary less. Query expansion that leads to less variation might be seen in two ways. First, such query expansion should improve performance in terms of precision and recall by bringing to the fore relevant documents that do not include the same terminology as the original query. Second, such query expansion should make document order in the relevant-document subsequences less dependent on the particular terminology used in the query. In terms of a dissimilarity measure that reflects the ordering of relevant documents, query expansion should reduce the dissimilarities among responses to alternative statements of the same information need. In this paper, we pursue this second manifestation.

Discussion of the approach introduced in this paper begins in the next section with specification of the dissimilarity measure. Analysis of the dissimilarities thus computed requires multidimensional scaling for graphical presentation as illustrated in Section 3. Finally, future work needed to take full advantage of the approach is discussed in Section 4.

2. DISSIMILARITY MEASURE

The dissimilarity measure used in this paper leads to fresh insights from the TREC 1000-document lists. Since two such lists can be compared in many ways, there are alternative measures. It is a contribution that we have found an effective measure although it may not be the most effective. We begin this section by specifying our dissimilarity measure and then review alternatives.

For each TREC topic, there is available a set of documents that assessors have determined to be relevant to the topic. For the topic under consideration, let there be n_R documents in this set, and let these documents be indexed by $i = 1,..., n_R$. In a TREC 1000-document list, let n be the number of relevant documents returned, and if relevant document i is returned, let r_i $(1 \le r_i \le 1000)$ denote its position in the list.

As the basis of our dissimilarity measure, we let R_i denote the position of document i in what is left of the list when the irrelevant documents have been removed. In other words, we let R_i denote the rank of r_i among the positions of the relevant documents, $r_1,..., r_n$. Thus, if $r_i = \min(r_1,..., r_n)$, then $R_i = 1$, that is , document i is returned first among the relevant documents. To the relevant documents not returned, we assign the same R_i, the average of ranks $n + 1$ to n_R. Thus, if relevant document i is not returned, we let $R_i = (n_R + n + 1)/2$. Our dissimilarity measure is based on the R_i thus defined. Note that irrelevant documents positioned in different ways in a 1000-document list can lead to the same $R_1,..., R_{n_R}$. The irrelevant documents influence our dissimilarity measure only through n, the number of relevant documents returned.

Our dissimilarity measure is obtained from Spearman's coefficient of rank correlation (Gibbons, 1985). Consider two 1000-document lists with $n^{(p)}$ relevant documents returned in the first and $n^{(q)}$ in the second and with relevant document return order $R_1^{(p)},..., R_{n_R}^{(p)}$ for the first and with $R_1^{(q)},..., R_{n_R}^{(q)}$ for the second. Spearman's coefficient of rank correlation adjusted for the relevant documents not returned is given by

$$s_{pq} = \frac{n_R^3 - n_R - 6\sum_{i=1}^{n_R}(R_i^{(p)} - R_i^{(q)})^2 - (U^{(p)} + U^{(q)})/2}{\sqrt{[n_R^3 - n_R - U^{(p)}][n_R^3 - n_R - U^{(q)}]}},$$

where

$$U^{(p)} = (n_R - n^{(p)})^3 - (n_R - n^{(p)})$$

$$U^{(q)} = (n_R - n^{(q)})^3 - (n_R - n^{(q)}).$$

Converting s_{pq}, which is a similarity measure, to our dissimilarity measure, we obtain

$$\delta_{pq} = \sqrt{1 - s_{pq}}.$$

To develop an understanding of this dissimilarity measure, one might consider the case in which all the relevant documents are returned in both lists. In this case, s_{pq} is just the product-moment correlation coefficient computed from the ranks $R_1^{(p)},\ldots, R_{n_R}^{(p)}$ and $R_1^{(q)},\ldots, R_{n_R}^{(q)}$. Moreover, δ_{pq} is proportional to

$$\sqrt{\sum_{i=1}^{n_R} (R_i^{(p)} - R_i^{(q)})^2}.$$

The contribution of relevant document i to this quantity is the difference $R_i^{(p)} - R_i^{(q)}$, the difference between the two lists in the relevant-document rank of document i. Thus, δ_{pq} measures dissimilarity in terms of the relevant documents at the fore in each list.

To put our dissimilarity measure in context, we consider its relation to performance measures and to other dissimilarity measures. Because the dissimilarity between two lists can be defined as the absolute value of the difference between the performance measures for the two lists, one can see how to turn a performance measure approach into a dissimilarity approach. The reverse is not generally possible in the sense that one usually cannot find a univariate performance measure that gives the dissimilarities among several lists. This is not surprising since one would expect that description of the differences among 1000-document lists would require many dimensions.

A popular performance measure is average precision. One can calculate it by first sorting the relevant document positions $r_1,\ldots, r_n$ to obtain $r_{(1)},\ldots, r_{(n)}$, where $r_{(1)} \le r_{(2)} \le \ldots \le r_{(n)}$ and then computing

$$\frac{1}{n_R} \sum_{i=1}^{n} \frac{i}{r_{(i)}}.$$

Note first that this performance measure involves only the distinction between relevant and irrelevant documents and nothing else derived from the document identifiers. Other performance measures based one way or another on precision and recall have this same property. It is moreover true that because the documents in the collection are not graded according to relevance, there is no way to define a performance measure that involves more than the relevant-irrelevant distinction. This immediately shows that our dissimilarity measure involves a novel aspect of system responses. Performance measures involving precision and recall are generally measures of how well irrelevant documents are rejected. On the other hand, our dissimilarity measure makes different use of the document identifiers and thereby opens up the possibility of new insights from TREC data.

There are dissimilarity measures other than the one specified in this paper and those based on performance measures. One might invent a dissimilarity measure that reflects the difference

between the irrelevant documents in two lists. One might compare the return order of relevant documents by computing Kendall's tau instead of Spearman's coefficient of rank correlation. It is possible that use of a few more dissimilarity measures would give further insight into TREC data.

3. DISPLAY BY MULTIDIMENSIONAL SCALING

Say that one wants to compare a group of system responses (1000-document lists) and that one computes a dissimilarity for each pair in the group and thus the dissimilarity matrix for the group. As argued in Section 2, this might lead to insights that cannot be obtained from any performance measure. However, looking at the dissimilarity matrix is unlikely to produce much insight. Rather, insight can be obtained from a dissimilarity matrix through multidimensional scaling (Cox and Cox, 1994; Kruskal and Wish, 1978; Rorvig, 1999). This technique produces points on a plane, one point for each system response, arranged so that the Euclidean distances between the points approximate the dissimilarities. Thus, one obtains the system responses laid out in a two-dimensional configuration with more dissimilar responses farther apart. The configuration can then be further interpreted. The multidimensional scaling algorithm we use is Kruskal's isotonic multidimensional scaling, which is named "isoMDS" by Venables and Ripley (1999).

In this paper, we consider four topics that are interesting in themselves and illustrate the kinds of results one can obtain. For each topic, we compare six systems in terms of their responses to twenty queries. Thus, for each topic, we apply multidimensional scaling to a 120 by 120 dissimilarity matrix. The six systems are "IN7a," which is a version of the INQUERY system from the University of Massachusetts; "Saba," which is a version of the SMART system from SabIR Research; "humA," which is a Hummingbird system; "ok9u," which is a version of the Okapi system from Microsoft; "IN7e," which is another version of the INQUERY system; and "Sabe," which is another version of the SMART system. The first four systems do not employ query expansion whereas the last two do. (Further description of these systems is found elsewhere in this publication.) The twenty queries for each topic are, after removal of duplicates, the best performing according to a criterion based on recall-at-1000 values (given by n/n_R) for the six systems. Our criterion is a weighted combination of the six recall values with weights for the expansion systems twice as large because the number of expansion systems considered is half the number of non-expansion systems. The use of exactly this criterion is not essential to the results in this paper.

Multidimensional scaling gives a plot with a point for each query-system combination. As our plotting symbols, we combine a query symbol with a system symbol. The query symbols for the 21 queries carried over from TREC-8 are A, B, ..., U, respectively; and the query symbols for the 22 queries new in TREC-9 are v, w, a, b, ..., t, respectively. The system symbols are 1, 2, 3, 4, *, and #, respectively.

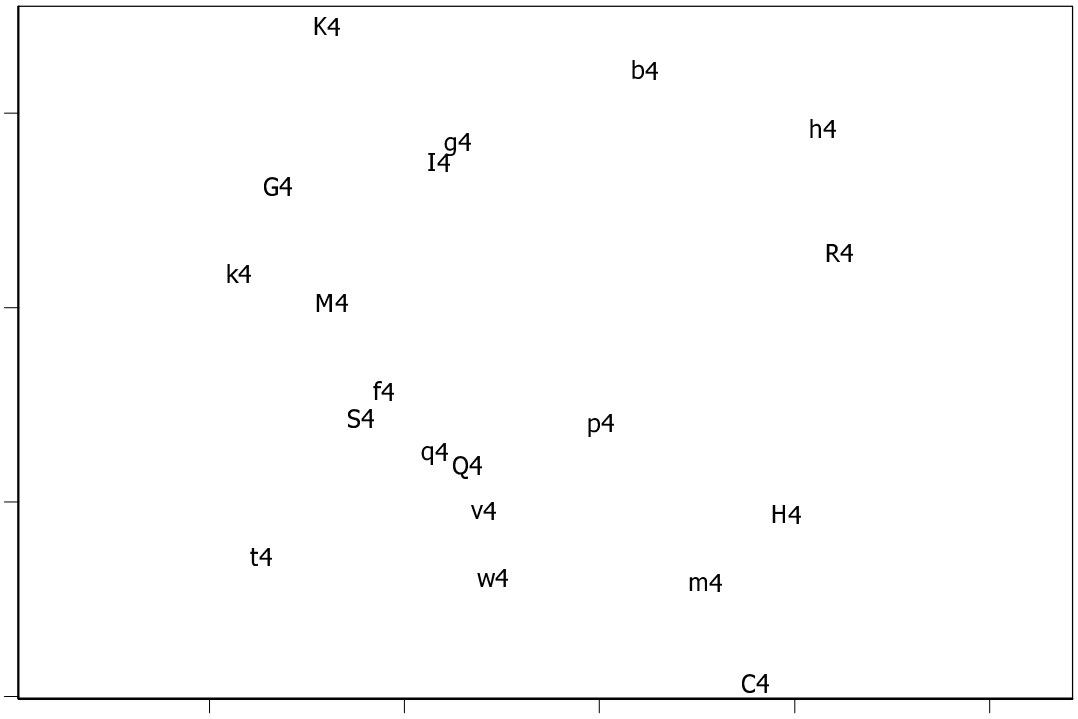

Figure 1. System “ok9u” Points for Topic 100.

Our first example is multidimensional scaling for topic 100. We do not begin by showing all 120 responses, however. In Figure 1, we show only the 20 points that correspond to the system “ok9u.” We have blanked out the points corresponding to the other systems. Each point in this figure corresponds to a query. The meaning in this figure is obtained from relative distances among points. For example, we see that R4 is closer to h4 than to t4. In other words, R4 and h4 are less dissimilar than R4 and t4 or h4 and t4. Looking at the queries, we see that this is reasonable since R4 is “cocom control export,” h4 is “America enforcing the terms of the COCOM agreement,” and t4 is “policy, regulation or control of high technology transfer.” Essentially, R4 and h4 are closer together because they share the term “cocom.” Because Figure 1 shows only relative distances, the configuration of points can be shifted, scaled up or down by the same amount on each axis, and rotated without affecting the meaning. This is the reason why no values are attached to the axis tick marks.

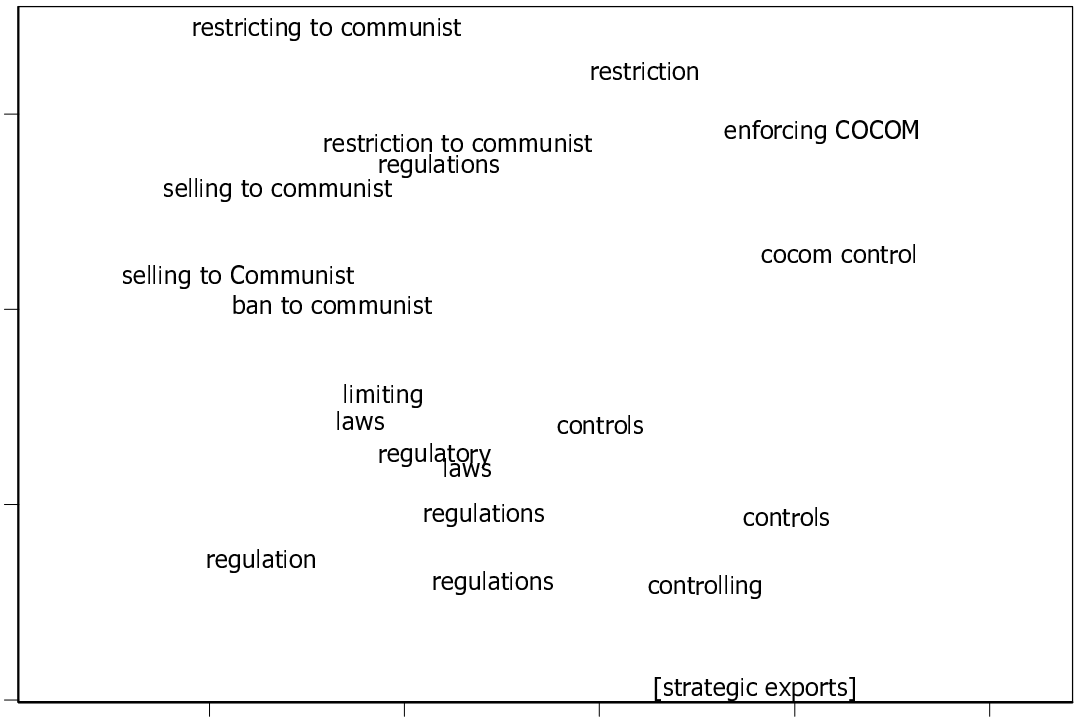

Figure 2. Selected Query Words at the "ok9u" Points for Topic 100.

Figure 2 is a somewhat subjective association of query texts with all the points in Figure 1. Because space on the figure prohibits printing the entire texts, we have selected a few words from each query. The word "export" appears in all the queries so we have omitted this word except in the one case in which the entire query is "strategic exports." What we see in Figure 2 are two axes that give meaning to the configuration. Horizontally, we see that the queries vary from reference to laws and regulations on the left to reference to control on the right. Vertically, we see that the queries vary from reference to general threats on the bottom to communist threats on the top. Thus, we see the variations in query wording that lead to the major differences in the response of the "ok9u" system.

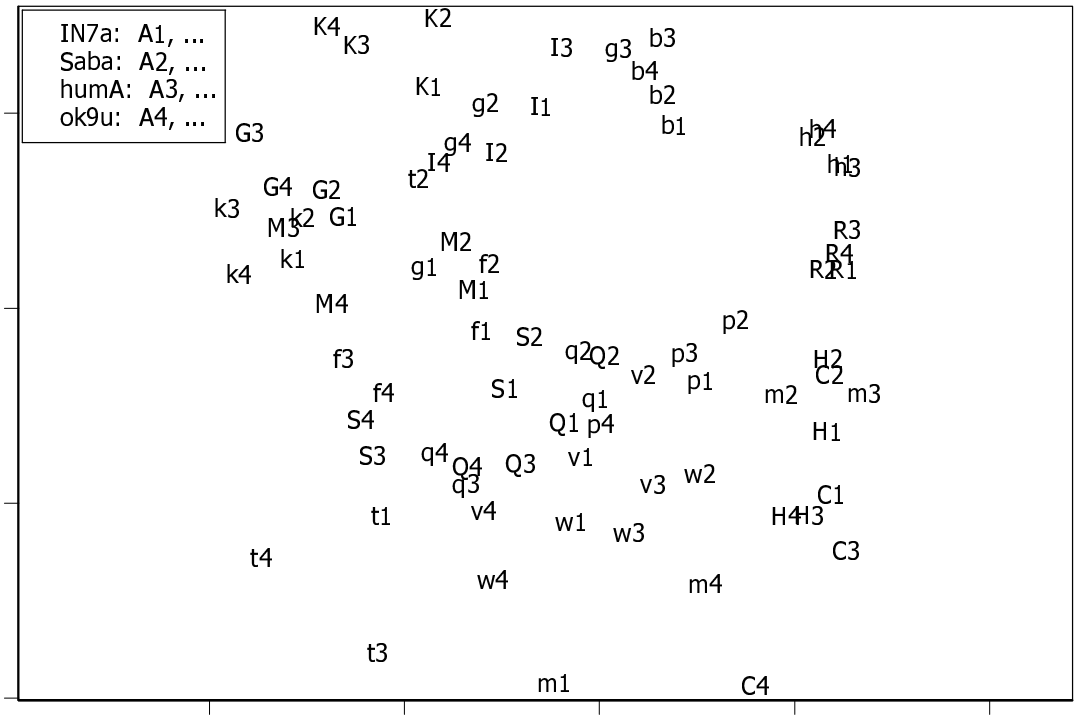

Figure 3. Topic 100 Points for the Non-Expansion Systems.

For topic 100, these query-wording axes hold for all the systems that do not employ query expansion. Figure 3 shows this. For a particular letter, the points with numbers 1, 2, 3, and 4 are generally close together. There are exceptions such as the distance t2 is from t1, t3, and t4. Nevertheless, if we were to associate query text with points for systems "IN7a," "Saba," or "humA," the resulting figures would be much like Figure 3. Thus, for this topic, the queries provide system-independent meaning to the space created by multidimensional scaling.

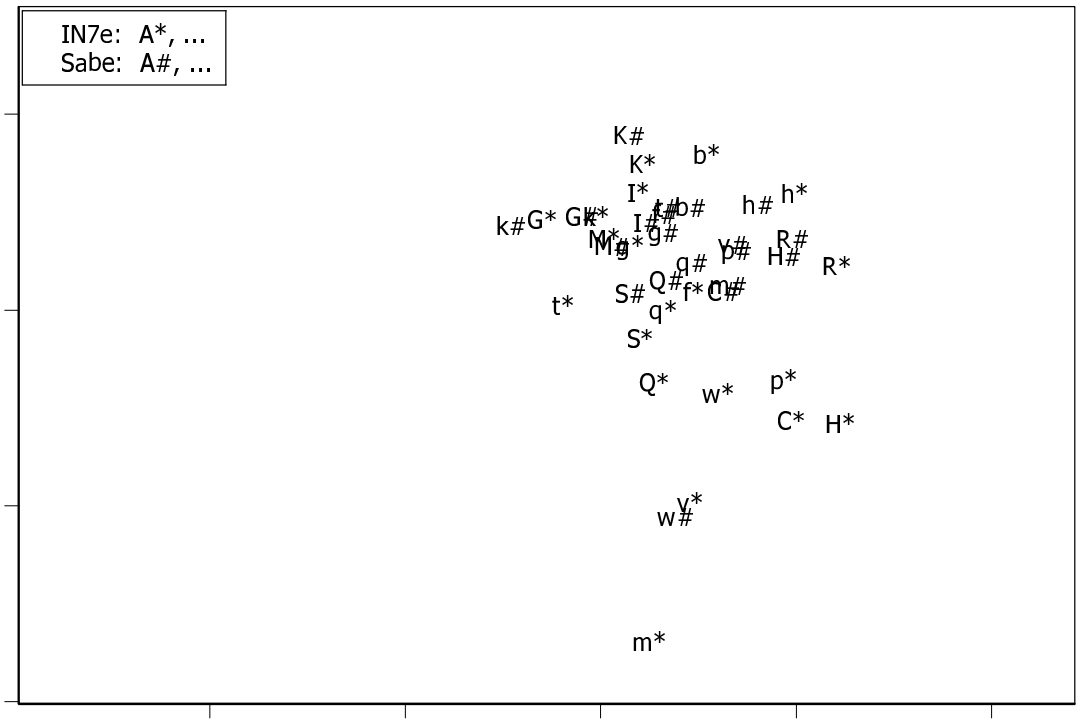

Figure 4. Topic 100 Points for Systems with Query Expansion.

Figure 4 shows that for Topic 100, query expansion is effective in the sense that it reduces the variation in system response due to query-to-query variation. In comparison of Figure 4 with Figure 3 (which both have the same scale), we see that the scatter in the responses of systems "IN7e" and "Sabe" is much less. Thus, query expansion as incorporated in these two systems makes the system response less dependent on the particular words chosen to express the topic, the need for information. In particular, there is less dependence on whether the query uses "control" or "regulation" and whether or not the query includes the term "communist."

One might question the point in Figure 4 labeled "m*." The response to this query is apart from the other responses in this figure. This point is the response of the system "IN7e" to the query "U.S.'s controlling of international exports." Given this query, this system was unable to retrieve documents that were retrieved by system "Sabe" with this query and documents retrieved by both systems with other queries. Noting an occurrence such as this could lead to insight into how a system can be improved.

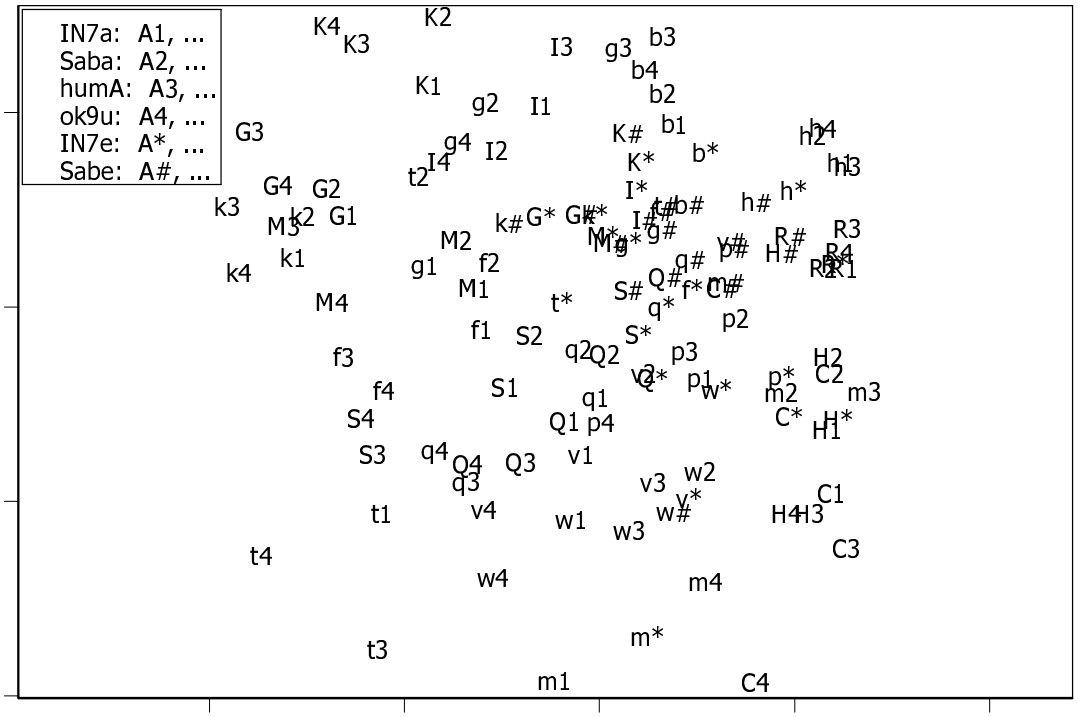

Figure 5. Multidimensional Scaling for Topic 100, All Points.

Figure 5 shows the configuration given by all 120 responses. Note that Figures 1-4 are all based on this configuration, that is, Figures 1-4 each exhibits only some of the 120 points but these at the locations shown in Figure 5. This figure is the one that actually gives the multidimensional scaling result that we use to interpret Topic 100. This figure gives an overview of all six systems. With the introduction provided by Figures 1-4, this overview might be helpful. As a place to start the analysis of a topic, a figure such as Figure 5 may require considerable effort before a reasonably complete interpretation can be obtained.

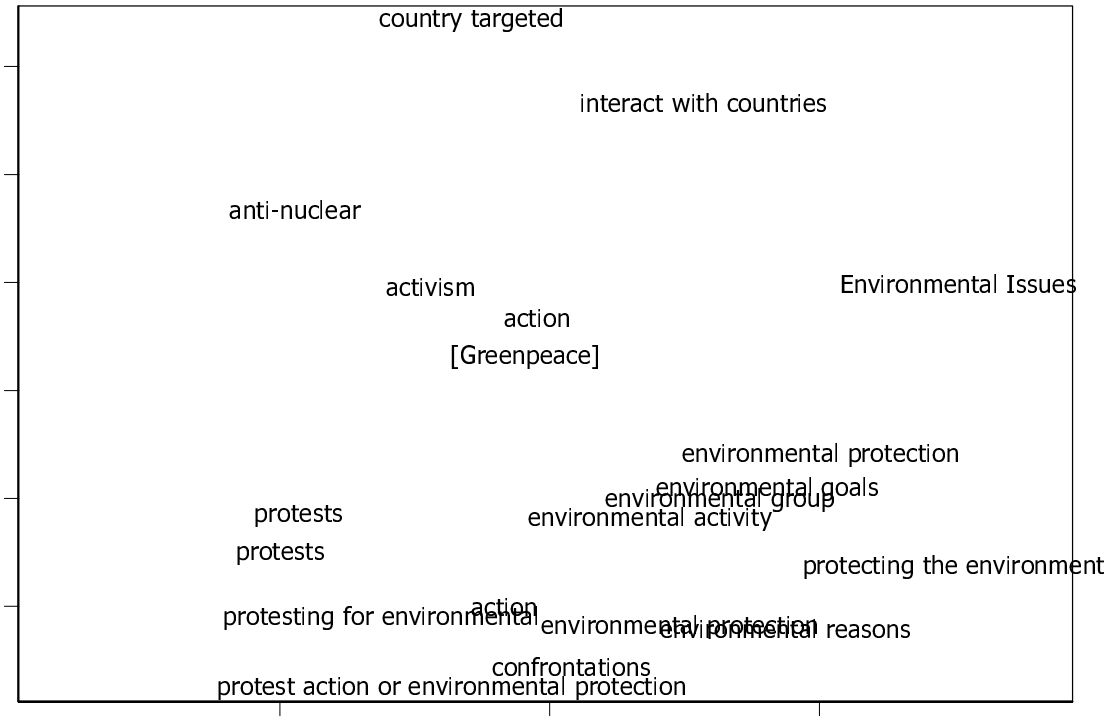

Figure 6. Selected Query Words at “ok9u” Points for Topic 78.

In our presentation of Topic 78, we begin with terms from the queries positioned at the points given by the system “ok9u” as shown in Figure 6. Because the 20 queries we consider all include the term “Greenpeace,” we have omitted this term except in the case of the query that consists of the single word “Greenpeace.” One way to interpret Figure 6 is to regard the horizontal axis as distinguishing Greenpeace regarded as a protest organization on the left and Greenpeace regarded as an environmental organization on the right. A proper interpretation of the vertical axis is less clear. Maybe the vertical axis distinguishes queries that refer to the actions Greenpeace takes from queries that refer to the targets of Greenpeace actions.

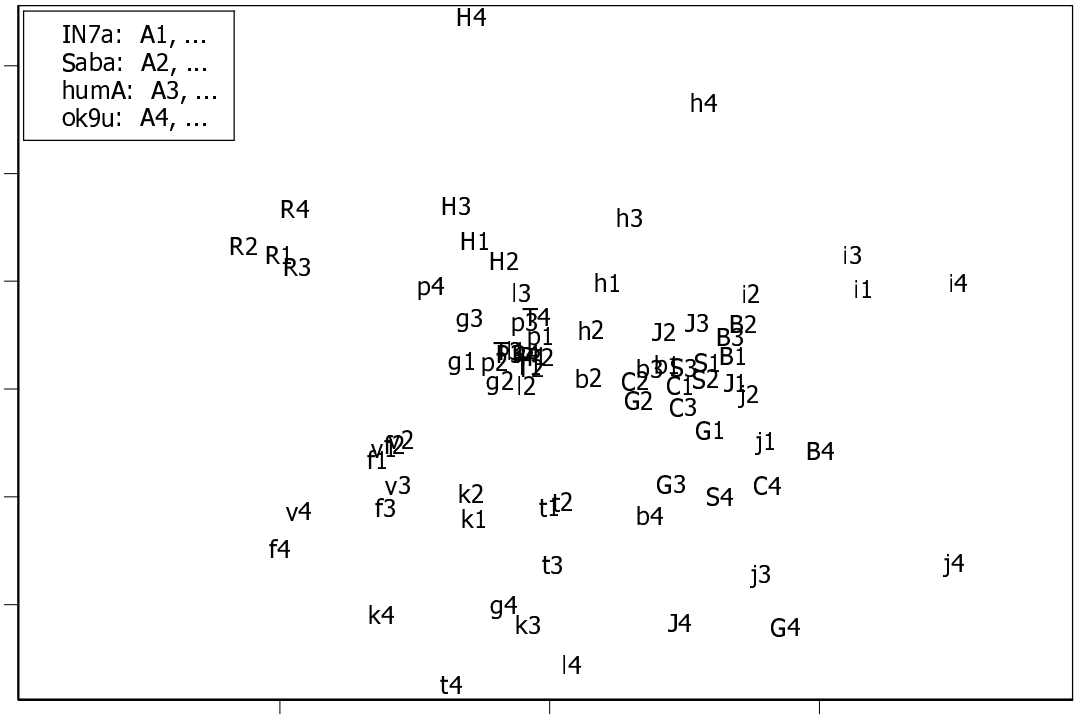

Figure 7. Topic 78 Points for Non-Expansion Systems.

Figure 7 shows that the most influential query terms affect the other non-expansion systems, “IN7a,” “Saba,” and “humA,” as they do “ok9u.” Generally, for each query, the four points for these four two systems lie close to each other. There are some exceptions such as the points g4, l4, and J4. Nonetheless, variation over the space portrayed by multidimensional scaling has meaning beyond the response of a particular system. This is the same observation that we made about topic 100 in conjunction with Figure 3.

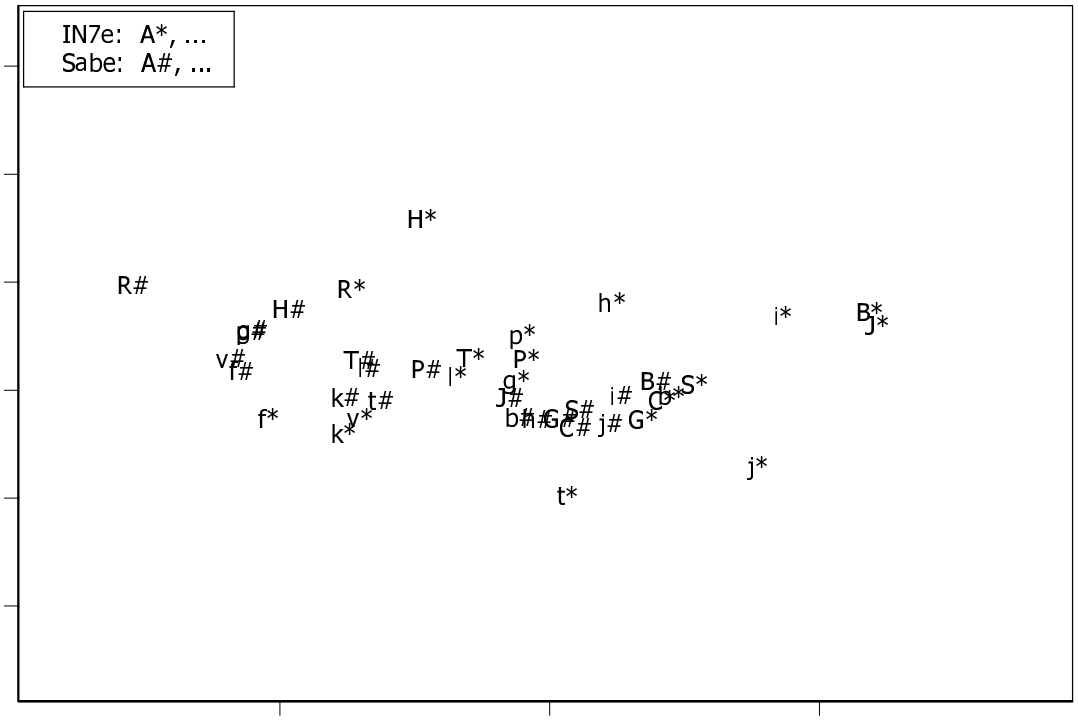

Figure 8. Topic 78 Points for Systems with Query Expansion.

Figure 8 shows the part of the configuration for Topic 78 produced by the systems with query expansion. Here, compared to Figure 7, we see less scatter in the vertical direction. Variation in the horizontal direction seems to have two properties. First, we see that for a query, the point for the system “Sabe” generally lies to the left of the point for the system “IN7e.” One might say that the system “Sabe” tends to regard Greenpeace as a protest organization and that the system “IN7e" tends to regard Greenpeace as an environmental organization. What characteristics of the query expansion algorithms this reflects is an interesting question. Second, it seems that each system reduces the scatter in the horizontal direction but that this reduction is not toward the same point on the “protest” - “environmental” axis.

Because what can be learned has largely been shown in Figures 7 and 8, we omit the figure showing the entire configuration for Topic 78. This omitted figure does provide a better basis than Figures 7 and 8 for observing that the tendency of “Sabe” to regard Greenpeace as a protest organization is true with respect to the non-expansion systems as well as “IN7e.”

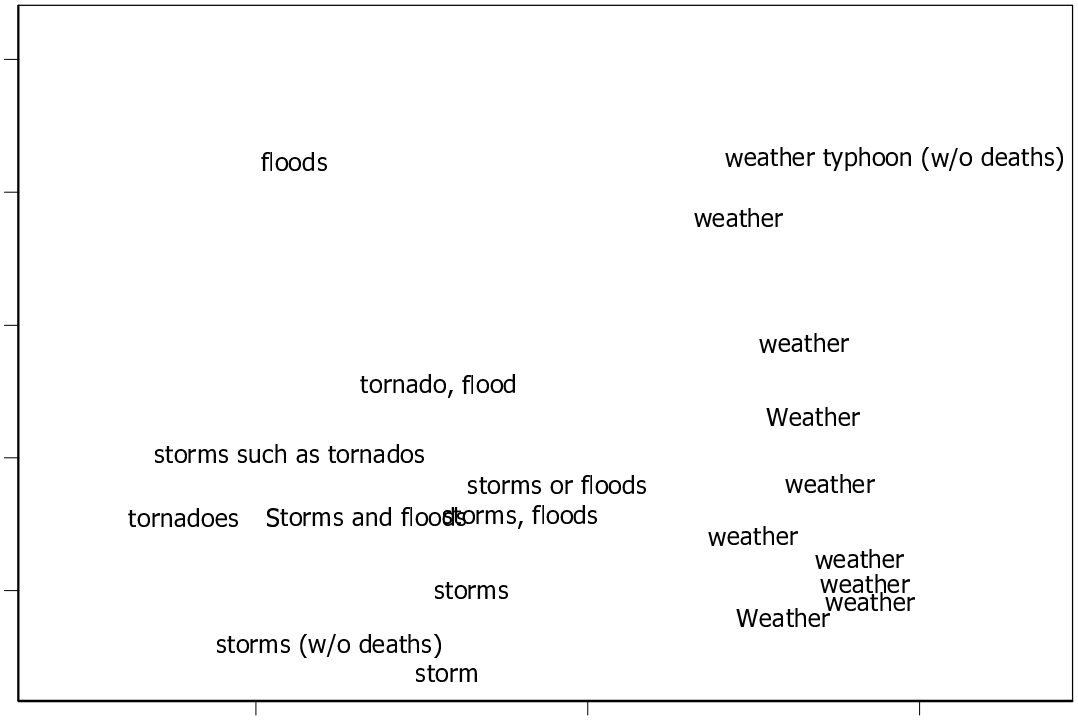

Figure 9. Selected Query Words at “ok9u” Points for Topic 59.

In our presentation of Topic 59, we again begin with terms from the queries positioned at the points given by the system “ok9u.” These terms, which are shown in Figure 9, do not include the word “deaths” because this term occurs in almost all queries. Rather, we have indicated queries that do not include the word “deaths.” In the horizontal direction, Figure 9 shows a clear separation between “storms” and “weather.” The phrase “storm-related deaths” is equivalent to the phrase “weather-related deaths.” Yet, as we will see, all of the systems respond as though these two phrases have different meaning.

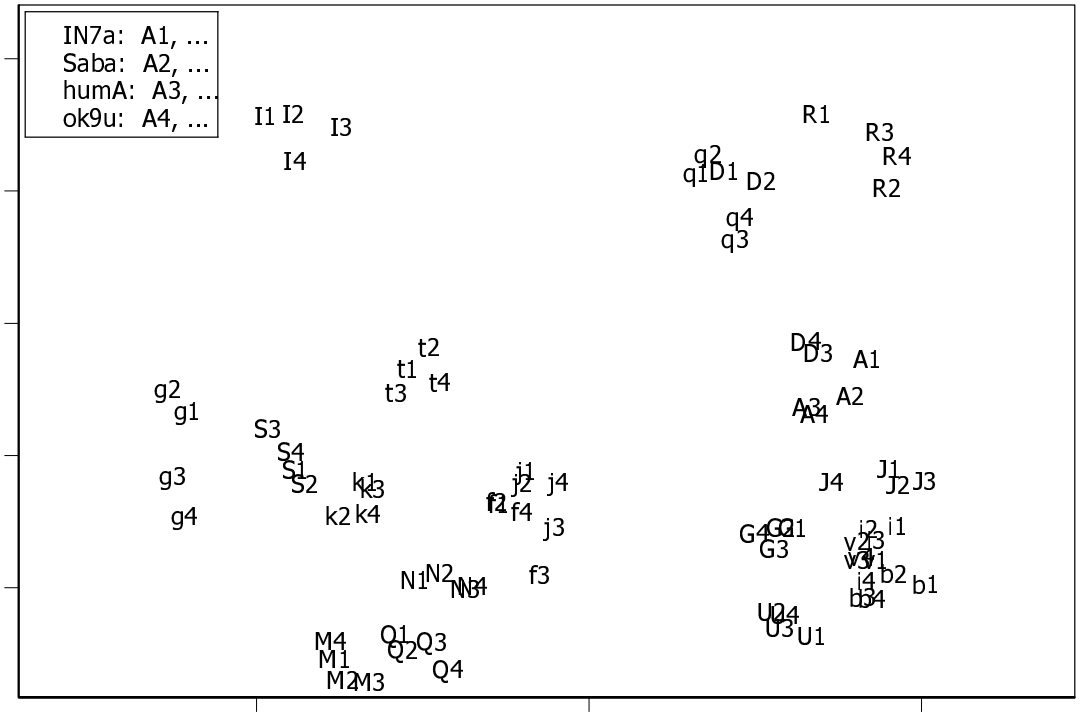

Figure 10. Topic 59 Points for Non-Expansion Systems.

Figure 10 shows the part of the configuration given by the non-expansion systems. These four systems respond similarly to each query. The scatter in this figure is largely query related, not system related. We see that the relative locations of query terms shown in Figure 9 apply to the other non-expansion systems as well.

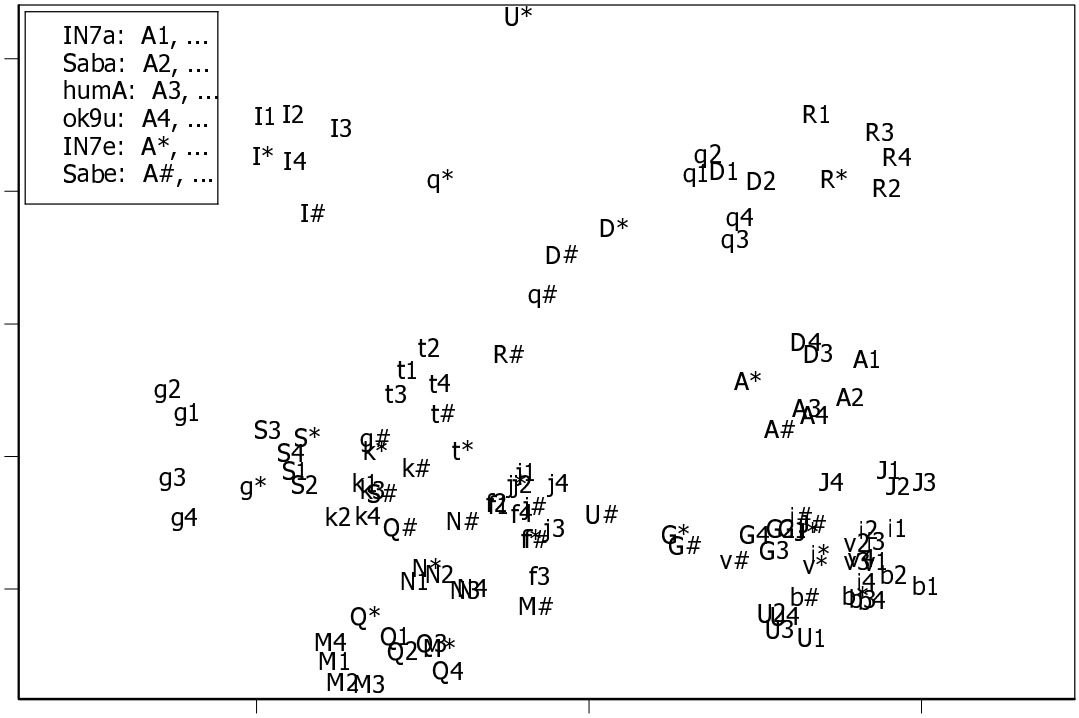

Figure 11. Multidimensional Scaling for Topic 59, All Points.

Figure 11 shows all 120 points in the topic 59 configuration. We see that although the query expansion systems move points associated with some queries, neither expansion system offers much reduction in the query-to-query scatter. Moreover, the “storm-related” - “weather-related” dichotomy also exists for these systems. One could draw a line just to the left of points U#, R#, and q* which would separate these two categories for all six systems. Thus, Topic 59 provides an example of failure of query expansion, failure to associate “storm-related” and “weather-related.”

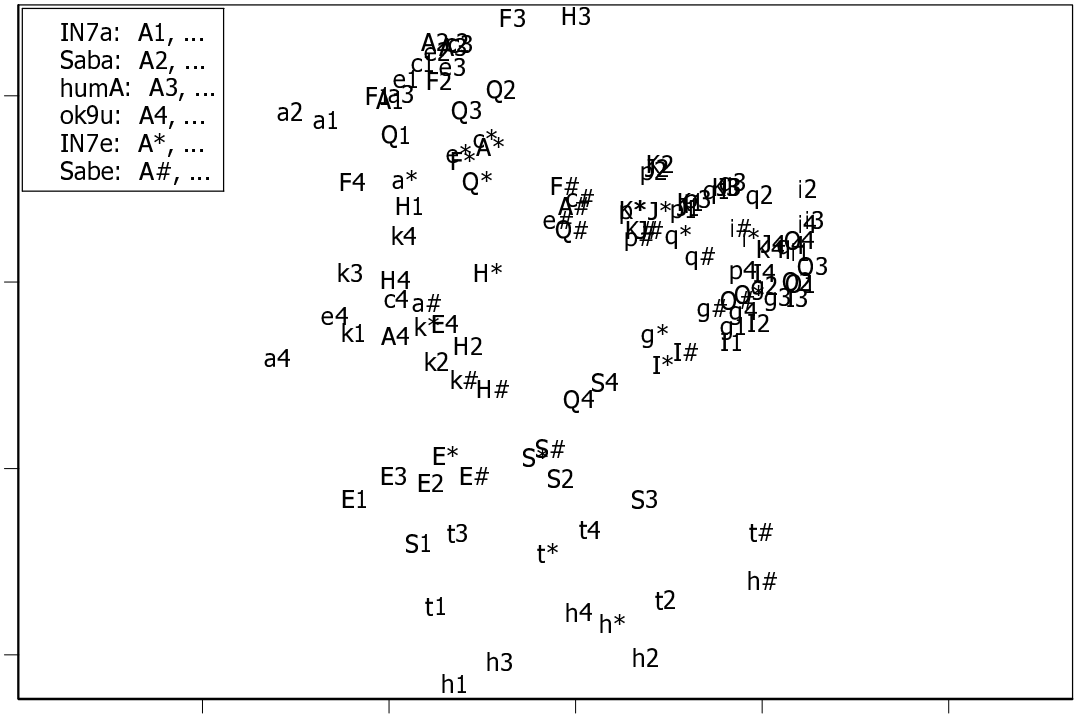

Figure 12. Multidimensional Scaling for Topic 86, All Points.

Results for Topic 86 cannot be summarized in the same way as the other topics discussed above. We begin with the entire configuration, which is shown in Figure 12. Study of this figure shows that whereas for topic 59, query differences are generally much greater than system differences, the situation for topic 86 is less clear. To show this, we compare the responses of “ok9u” with those of “Sabe.”

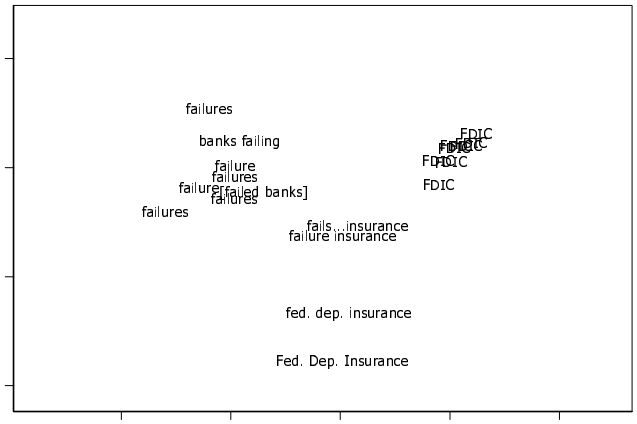

Figure 13. Selected Query Words at "ok9u" Points for Topic 86.

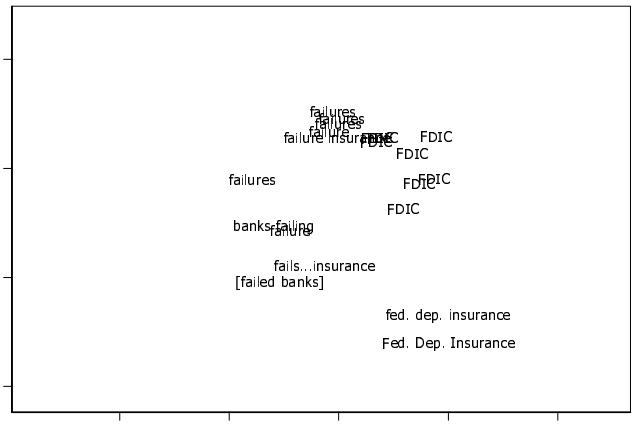

Figure 14. Selected Query Words at "Sabe" Points for Topic 86.

Figure 13 shows terms from each query at the "ok9u" locations. Consider the queries in three groups, those containing the term "FDIC," those containing the phrase "Federal Deposit Insurance Corporation," and those that do not refer to the FDIC in either way. All the queries in this third group contain the word "bank." Figure 13 shows that "ok9u" does not associate "FDIC" with the phrase spelled out although it should. In fact, none of the six systems do. The queries with "FDIC" form a tight cluster. The other queries are more spread out.

Figure 14 shows the same query terms as Figure 13 but at the "Sabe" locations. We see that "Sabe," an expansion system, brought some but not all of the queries in the third group closer to the "FDIC" queries. One would guess that this is caused by other terms in the queries but which terms is not clear. For topic 86, variation across the space defined by multidimensional scaling involves both query effects and system effects. Thus, interpretation of the configuration for topic 86 is difficult.

It is possible to summarize results from these four topics. For topic 100, query expansion reduces the variation due to restatement of the topic as one would hope. For topic 78, query expansion also reduces the variation due to restatement but the two expansion systems do this differently. For topic 59, query expansion does not recognize one equivalence in the query statements, the equivalence between "storm-related" and "weather-related." For topic 86, query expansion fails in a more complex way.

4. CONCLUSIONS

The claim in this paper is that beyond differentiation of relevant and irrelevant documents, more insight can be obtained from the document identifiers that are part of the TREC system responses. In particular, we consider the return order of relevant documents compared by means of Spearman's coefficient of rank correlation. We have supported our claim by showing that for specific topics, the return order of relevant documents can help us understand the difference between systems with and without query expansion. This paper opens the door to many more possibilities for insights.

Our claim is not that a dissimilarity matrix computed from the return order of relevant documents is a substitute for a performance measure of the precision and recall variety. One important way of going beyond the analysis in this paper is extension to an analysis of both the dissimilarities in this paper and a selected performance measure such as average precision. Computing a new dissimilarity measure from the two is a possibility but perhaps not the best idea. Rather, one should realize that there may be topics for which the performance measure does little to distinguish the system returns but the return order of relevant documents is much more informative. There may be topics for which the converse is true. These are the topics for which further insight can be obtained by considering in addition to the usual performance measures, the return order of relevant documents.

One possibility would be to compute the performance difference between expansion and non-expansion systems for all the topics and use this series of numbers to pick topics to be looked at in terms of the return order of relevant documents. Such an approach seems necessary because looking individually at all 50 topics seems overwhelming in light of the four or five figures that each topic requires for interpretation. Thus, the return order of relevant documents would be the basis for analysis of the failures of query expansion.

One would like to summarize what is shown by our return order of relevant documents over all 50 topics. This is not as easy as with a performance measure that can be averaged over the topics. One can however, think of quantifying what is observed in the topics discussed above. In the case of Topic 100 in particular, one can think of a measure of scatter that would one could use to evaluate the effectiveness of query expansion. One could then compute such a measure

for each topic and use it to summarize over topics. One could then rank topics by the effectiveness of query expansion and investigate a sampling of topics in detail. Such an investigation could be the next step.

REFERENCES

D. Banks, P. Over, and N. Zhang (1999). "Blind Men and Elephants: Six Approaches to TREC Data," *Information Retrieval* 1, 7-34.

M. W. Berry and M. Browne (1999). *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

C. Buckley and J. Walz (2000). "The TREC-8 Query Track," In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8),*" pages 65-75, NIST Special Publication 500-246, Washington, DC: US Government Printing Office.

T. F. Cox and M. A. A. Cox (1994). *Multidimensional Scaling*, London: Chapman & Hall.

J. D. Gibbons (1985). *Nonparametric Statistical Inference*, New York: Marcel Dekker.

J. B. Kruskal and M. Wish (1978). *Multidimensional Scaling*, Newbury Park, CA: SAGE Publications.

M. Rorvig (1999). "Images of Similarity: A Visual Exploration of Optimal Similarity Metrics and Scaling Properties of TREC Topic-Document Sets," *Journal of the American Society for Information Science*, 50, 639-651.

W. N. Venables and B. D. Ripley (1999). *Modern Applied Statistics with S-PLUS, Third Edition*, New York: Springer-Verlag.