Context-Based Statistical Sub-Spaces

TREC-6 Notebook Paper

Gregory B. Newby
University of North Carolina at Chapel Hill[*]

ABSTRACT

The technique described in this paper is similar to latent semantic indexing (LSI), although with some variation. Whereas LSI operates by performing a singular value decomposition (SVD) on a large term by document matrix of co-occurrence scores, the technique here operates by identifying eigenvectors and eigenvalues of a term by term matrix of correlation scores (derived from co-occurrence scores). The technique of identifying eigenvectors and eigenvalues from a correlation matrix is known as principal components analysis (PCA). Variations from the previous year's TREC work include work using sub-documents (paragraphs), and working with small sub-matrices consisting only of terms in a query, rather than working with all terms from the collection.

INTRODUCTION

The approach to TREC-6 described in this paper is based on principal components analysis, in which a term co-occurrence matrix is used as a basis for generating an "information space." Rather than pursuing a model that applies a single information space for all queries, this year's TREC effort builds a custom information space for each query. This "context based" approach is intended to better distinguish among documents which possess terms from the query than an approach that simply folds all terms and queries into a much larger generic information space.

Retrieval from the space is typical of vector-space and related methods, in that queries are represented as pseudo-documents and then document similarity scores are computed between each document and the query/pseudo-document. The two main differences are that term and document vectors are not normalized, and that the geometric distance measure is used, rather than a cosine, dot product, or other measure.

The space-building process is as follows:

1. Identify a list of "good" words (words of interest) for a collection.
2. Count the co-occurrence of each good word with other good words across all documents in the collection. This is the symmetric term by term co-occurrence matrix for the collection.

[*] Address: CB-3360 Manning Hall; Chapel Hill; NC; 27599-3360; USA. Email gbnewby@ils.unc.edu.

3. For a query, identify the query's unique "good" words. Build a co-occurrence matrix based on the query context by extracting only these "good" term pairs from the collection co-occurrence matrix.
4. Principal components analysis (PCA) is performed on the query correlation matrix (generated from the query co-occurrence matrix) to identify eigenvectors and eigenvalues. The eigenvectors are the basis of the multidimensional information space, in which each term has a known numeric (metric) relationship to all other terms. The eigenvectors are used as coordinates of the query words in a multidimensional space.
5. Documents are located at the geometric center of the terms they contain.
6. Queries are represented as pseudo-document by also locating them at the geometric center of the terms they contain.
7. Retrieval proceeds by determining which documents are closest to the query, with closer documents interpreted as "more relevant."

Three TREC-6 tasks were performed: the adhoc task, the routing task and the filtering task. All tasks were performed on the "Category A" dataset (the full dataset), and all used only "automatic" query construction and retrieval. Query expansion was not used, nor was term weighting.

Comparable techniques were used by the author in TREC-5 (Newby, 1996). Additional techniques employed beyond the TREC-5 methods are:

1. For all three tasks, a query-specific context subspace was generated. From all original "good" words, only those words in the query were used. This resulted in a far smaller co-occurrence matrix than for TREC-5, when the same 1900 by 1900 term co-occurrence matrix was used for all queries.
2. For the filtering task, sub-documents were used, rather than entire documents. Each sub-document was a paragraph from the original document.

The most important factor lacking in the approach described here appears to be term weighting. Due to a complete absence of term weighting, the results presented here indicate that documents were retrieved based on the presence of some query terms, but the terms were not necessarily conceptually "important" to the query. Specifically, query terms with high collection frequencies (*tf*) would result in retrieval of many documents with these high-frequency terms. Meanwhile, the more important query terms, which generally had lower *tf*'s, added relatively fewer documents to the retrieved set. Current efforts are being directed at identifying useful term weighting schemes. Because the term by term co-occurrence score, which is the basis for the technique here, is derived from two separate terms, it is not yet clear whether traditional *tf / idf* weighting will be appropriate.

The remainder of this document discusses the outcomes of each TREC-6 task. Following a section on visualization, a concluding section summarizes the work to date for TREC, and identifies the most important areas for continued development.

THE ROUTING TASK

The routing task was based on 50 queries with pre-existing relevance judgments. The data were new data from the FBIS (Foreign Broadcast Information Service), a total of 120,654 documents. Topics used were: 1 3 4 5 6 11 12 23 24 44 54 58 62 77 78 82 94 95 100 108 111 114 118 119 123 125 126 128 142 148 154 161 173 180 185 187 189 192 194 202 228 240 282 10001 10002 10003 and 10004.

21,494 “good” words were identified for this task. The words were derived from various lists of dictionary terms, places and proper names. In addition, all terms from the 50 queries were put on the “good” words list. 21,494 is a post-stemming count – a Porter stemming algorithm was adapted from Frakes & Baeza-Yates (1992). A stoplist based on the SMART team’s list was utilized (596 words).

Each of the 120,654 documents was analyzed for its contribution to a 21,494 by 21,494 co-occurrence matrix. For each document:

- All unique (stemmed) terms in the document were identified.
- Only those terms on the “good” words list were kept.
- For each term pair (**[N * (N-1)] / 2** pairs per document), the co-occurrence score in the full co-occurrence matrix was incremented by one.

The frequency of terms within documents (*df*) was not taken into account. The co-occurrence matrix thus counts the number of documents with each term pair, not the raw frequency of term pairs within all documents. The possible range of co-occurrence scores for the 230,996,018 ([21494^2] / 2) term pairs is 0 to 120,654 (the identity row was pre-defined as 1, in order to prevent zero variance for any rows in which all other scores were zeroes).

The number of pre-judged relevant documents from prior years for the 50 queries ranged from 18 to 2661, with a median of 131. This work did not take judgments of non-relevant documents from prior years into account.

The retrieval process was as described in the Introduction, above. Specifically:

1. 21,494 “good” stemmed words were identified by culling various word lists and pre-selecting all non stoplist terms from the 50 queries.
2. The 120,654 FBIS documents were used to build a 21,494 by 21,494 co-occurrence matrix.
3. A context co-occurrence matrix of *only those terms from the query* was extracted from the larger 21,494 by 21,494 co-occurrence matrix. Matrices for queries ranged in size from 3 by 3 to 106 by 106, with a median of 29 by 29.

4. Principal components analysis was performed on the context matrix, resulting in a multidimensional information space.
5. All FBIS documents with at least 25% of the query terms were located in the space. Documents that didn't meet the 25% cutoff were assumed to be non-relevant.
6. All pre-judged documents from prior TRECs were also located in the space.
7. Results were produced by retrieving the closest FBIS documents to the query and to the pre-judged documents.

Steps 6 and 7 were the points of departure from the adhoc task (below). Because no basis was developed for deciding which of the pre-judged documents was most important, or for determining a minimum cutoff value for closeness, a round-robin approach was used. For example, if 49 pre-judged documents were available from prior TRECs for a given query, these 49 plus the query were alternatively used to retrieve the next closest document (up to the desired set size of 1000 documents per query). In this case, the result would be the 20 closest documents to each of these 50 targets – but intermixed. This is not a very good criterion, but because the absolute metric values for each sub-matrix were different, there was no single cutoff value to use across queries.

The exceptions to this round-robin approach occurred when either there were more than 200 documents pre-judged as relevant, or when fewer than 1000 FBIS documents total met the 25% criteria. In the first case, time constraints prevented dealing with more than 200 pre-judged documents (this eliminated some pre-judged documents for almost half of the queries). In the second case, sometimes fewer than 1000 FBIS documents had 25% of the query terms, especially for long queries, so fewer than 1000 FBIS documents could be ranked for retrieval.

A second set of routing results was submitted (not for assessment) in which only those documents closest to the query were retrieved. In other words, in which none of the pre-judged documents were taken into account. This makes this set of results more like those of an adhoc task. The goal was to make some approximation to the prior year's TREC-5 effort, but with using context-specific subspaces rather than one larger space for all queries. In fact, results from this method do appear to be better: exact precision scores were higher, and the overall average percentage of relevant documents retrieved per query was higher. Further analysis will identify additional trends between the TREC-5 and TREC-6 results.

Across the pooled cohort group, an average of 146.2 (Standard Deviation = 180.9) relevant documents per query were found, with a range of 4 to 890. The first set of routing results had an average of 56.4 (SD 77.7), with a range of 0 to 388. The second set had an average of 18 (SD 24), with a range of 0 to 113.

Across all queries, the first set of routing results succeeded in identifying 36% of all pooled relevant documents for the query (SD 20%). The second set retrieved 11% (SD 9.6%). In both cases, the Pearson correlation between the number of relevant documents was strongly correlated ($r = .88$; $p < .0001$) with the number of pooled relevant documents.

Exact precision scores for the first set ranged from 0.0 to .52, with a median of .10. For the second set, scores ranged from 0.0 to .12, with a median of .01.

The overall interpretation of these results is that the routing approach described here is effective at "query by example," in the case of the first retrieved set, where existing relevant documents are used as surrogate queries, and similar documents are found. Further investigation may help to discover whether the main benefit of query by example is the similar lengths of surrogate query documents and the real documents (versus shorter "real" queries), or other factors.

Additional investigation should be made of the performance of the techniques described here for clustering previously judged documents. Measuring the presence and densities of relevant or non-relevant clusters could be extremely effective for identifying useful regions in the information space for the retrieval of new documents.

THE ADHOC TASK

The adhoc task was based on 50 queries numbered 301-350. The data for the task were from discs 4 and 5: a total of 555,871 documents.

Adhoc Document Counts

Congressional Record	CR	27716
Los Angeles Times	LA	131896
Foreign Broadcast Information Service	FBIS	130471
Federal Review	FR	55630
Financial Times	FT	210158

The adhoc task made use of full documents (not sub-documents). Two sets of results were submitted. The first was created the same way as the second routing set described above. That is, the query was located in the multidimensional subspace, and the 1000 closest adhoc documents to the query were retrieved. (If fewer than 1000 adhoc documents had 25% or more of the query terms, only those documents that made the cutoff were retrieved.)

A second non-assessed set of results utilized only the Description field of the query. This resulted in very short queries, ranging from 2 to 10 "good" terms each. Because no query expansion was applied, results were uneven. Query 316, for example, had the description "A look at the roots and prevalence of polygamy." Only *roots*, *prevalence* and *polygamy* were on the list of "good" words, but *polygamy* did not occur in any of the 555,871 documents! The same 25% cutoff value was employed for this results set.

Across all 50 queries, a pooled total of 72,270 documents were judged based on submissions from all TREC participants (53,650 unique documents). On average, 92.2 (Standard Deviation 103.1) relevant documents per query were identified, with a range from 3 to 474

The first set of results (utilizing full queries) yielded an average of 5.5 (SD 8.1) relevant documents per query, with a range from 0 to 35. A correlation of -.53 ($p < .0001$) between the total pooled relevant documents per query and the number of relevant documents from the first set was found. This indicates that the approach taken was less effective when many relevant documents were found by the TREC cohort. Or, inversely, that the approach was more effective for more "difficult" queries, in which fewer relevant documents were found by the cohort. However, the correlation was practically identical (-.52, $p < .0001$) for the pooled cohort group, indicating a likelihood that the system described here was typical in this regard.

A correlation of .80 ($p < .0001$) between the number of relevant documents in the first set and the total number of pooled relevant documents indicates, as expected, that the number of relevant documents in the set increases with the total number of relevant documents.

Recall and precision scores for this set were poor, with exact precision scores ranging from 0.0 to .08, with a median of 0.0. Overall, this set identified a per-query average of 5% of the pooled relevant documents, with a range from 0% to 24% (SD 5.6%).

The second set had somewhat better performance figures, but the soundest interpretation appears to be that the improvement was simply due to the greatly decreased number of query terms. By seeking a minimum of 25% of query terms per retrieved document (before locating the document in the information space and ranking it for retrieval), a somewhat better response set might be expected simply by this Boolean approximation to a first cut. However, greater than 1000 documents were generated by this cut for each query – indicating that even if the Boolean approximation increases effectiveness, the information space distance ranking could still be an important component of the improvement over the first set's results. Further investigation of the role of the Boolean approximation is being made by the author.

For the second set, the overall statistics for the number of pooled judgments and relevant documents found is the same as for the first set (both set's statistics were derived from the overall TREC cohort group). But the mean number of relevant documents per query in this set was 9.92 (SD 19.3), versus 5.5 for the first set. The range was from 0 to 125.

The correlation between the number of relevant documents in the second set and the total number of pooled relevant documents was .67 ($p < .0001$), while the correlation between the number of relevant documents found by the cohort group and the number found in this set was not significant (at $\alpha = .05$). These scores indicate that the approach taken here exhibited a different pattern, overall, than the cohort for the relation between the number of pooled relevant documents and the number of relevant documents retrieved for a given query. This pattern was also different than for the first retrieved set.

While the recall and precision scores for this second adhoc set were only slightly less poor than for the first set, the exact precision scores were somewhat improved with a range from 0.0 to .18, with a median of .01. Overall, this set identified a per-query average of 9.1% of the pooled relevant documents, with a range from 0% to 35.5%.

In summary, the adhoc results generated by the approach described here were not outstanding. Rather than attribute the enhanced performance of the second set (with short queries) to the information space approach, a simpler interpretation is that the 25% cutoff yielded a better sub-set of documents to locate in the information space, thereby yielding better results.

The overall pattern of adhoc results indicates the context-based information space approach may have some potential for usefulness, but not without additional work. The first set's pattern of retrieving 5% of relevant documents is only somewhat better than the 2% expected by chance through random selection. But when specific queries are examined, there appears to be more promise: some queries in both sets had reasonable performance measures (i.e., close to or above the median). Based on the relatively superior performance of the second set, it appears wise to investigate methods for term weighting or automatic query processing (expansion or term removal) in order to maximize the effectiveness of the set of documents eligible for ranking and retrieval.

THE FILTERING TASK

The filtering task was based on the same 50 queries and data as the routing task (120,654 FBIS documents). These documents were broken into sub-documents by inserting paragraph tags into the original documents whenever a line started with two spaces. (Unfortunately, a number of documents did not meet this standard, notably some transcriptions of interviews. These documents were treated as one long sub-document.)

This process resulted in a total of 1,909,729 sub-documents. Paragraphs with no "good" words were ignored and not counted. The number of sub-documents per document ranged from 1 to 4726.

The rationale for working with sub-documents is twofold. First is an effort to decrease the size of "documents" (that is, sub-documents) to be retrieved. Because the size range for full documents is great, this could lead to less variance in the size ranges for sub-documents – hopefully preventing an uneven likelihood of longer documents being retrieved versus shorter documents. The second component of the rationale is to simply investigate the applicability of the multidimensional information space approach for identifying useful sub-documents.

Otherwise, the retrieval process was identical to the process for the routing task, but with differences as to the final selection criteria for which documents were included in the retrieved set. Unlike the routing and adhoc tasks, in which a ranked set of retrieved documents is presented, the filtering task documents were not ranked. Instead, a binary relevance judgment about whether to accept or reject each FBIS document was to be made independently of judgments for other documents.

Multiple evaluation criteria, F1 and F2, were applied to the filtering task. Essentially, F1 penalizes more heavily for non-relevant documents being retrieved (favoring high precision), while F2 had a lesser penalty for non-relevant documents but an added penalty for not retrieving known relevant documents (balancing high precision with high recall).

For the F1 evaluation criteria, a conservative distance value for a cutoff was needed. This cutoff would determine a hypersphere around each query within which all documents would be retrieved. Because the metric of each query space (that is, the range of eigenvectors) was different from other spaces, there was no acontextual method for determining a cutoff value. In other words, it was not possible to choose a value (such as "10 units") which would be suitable for all query information spaces.

The conservative cutoff was chosen as the *smallest* of the distances from all pre-judged documents to the query. Any FBIS document that was closer to the query than this distance was retrieved. This resulted in fewer than the maximum of 1000 documents being submitted for all but 5 queries (240, 194, 180, 173 and 125), with a low of 1 document being submitted (query 10002) and a median of 158. Unlike routing, distances of documents to previously judged relevant documents were not utilized. It will be interesting, in the future, to see the effects of taking the same exact approach for routing and filtering, with the only difference being the use of full documents versus sub-documents.

The retrieved set for function F1 was poor. Of the 47 queries, the lowest score was achieved on 22 queries (46.8%), and none of the scores exceeded the median. In all but three of the lowest achieved scores, the number of FBIS document retrieved exceeded the pooled total number of relevant documents for the FBIS collection. In all, the 17 queries for which *fewer* than the pooled total number of relevant document were retrieved had higher F1 scores than the 30 queries for which *greater* than the pooled number were retrieved. This points to a problem of not only having few relevant documents in the retrieved set, but also being unable to discard non-relevant documents.

The retrieved set for function F2 was considerably better than the set for F1. This set utilized a 25% cutoff for the minimum number of query terms per document. Because of the relatively high cutoff, fewer than 1000 documents were eligible for being assigned a location in the information space. The range of eligible documents was from 0 (queries 23 and 148) to 843 (query 202), with a median of 16.

In effect, the F2 set was produced by Boolean probabilistic methods, in that results were not ranked (a ranked set is mentioned below). One query achieved the highest F2 score (query 23, with no documents retrieved) and 8 queries (17%) were close to or above the median. Six queries (12%) achieved the lowest score. For 37 queries (79%), the number of documents retrieved was lower than the total number of relevant documents. The results of F1 and F2 are indicative of the merits of retrieving very small sets for the filtering task.

Two additional sets of filtering results were delivered based on ranking of documents, rather than on binary filtering. The first set, based on F1, simply included the closest 1000 documents per query. In other words, the same results set as would have been submitted for the routing task. In practice, the results were not the same, because the 25% cutoff was not applied. Instead, FBIS documents with *any number greater than 1* of the query terms were located in the information space and eligible for ranked retrieval.

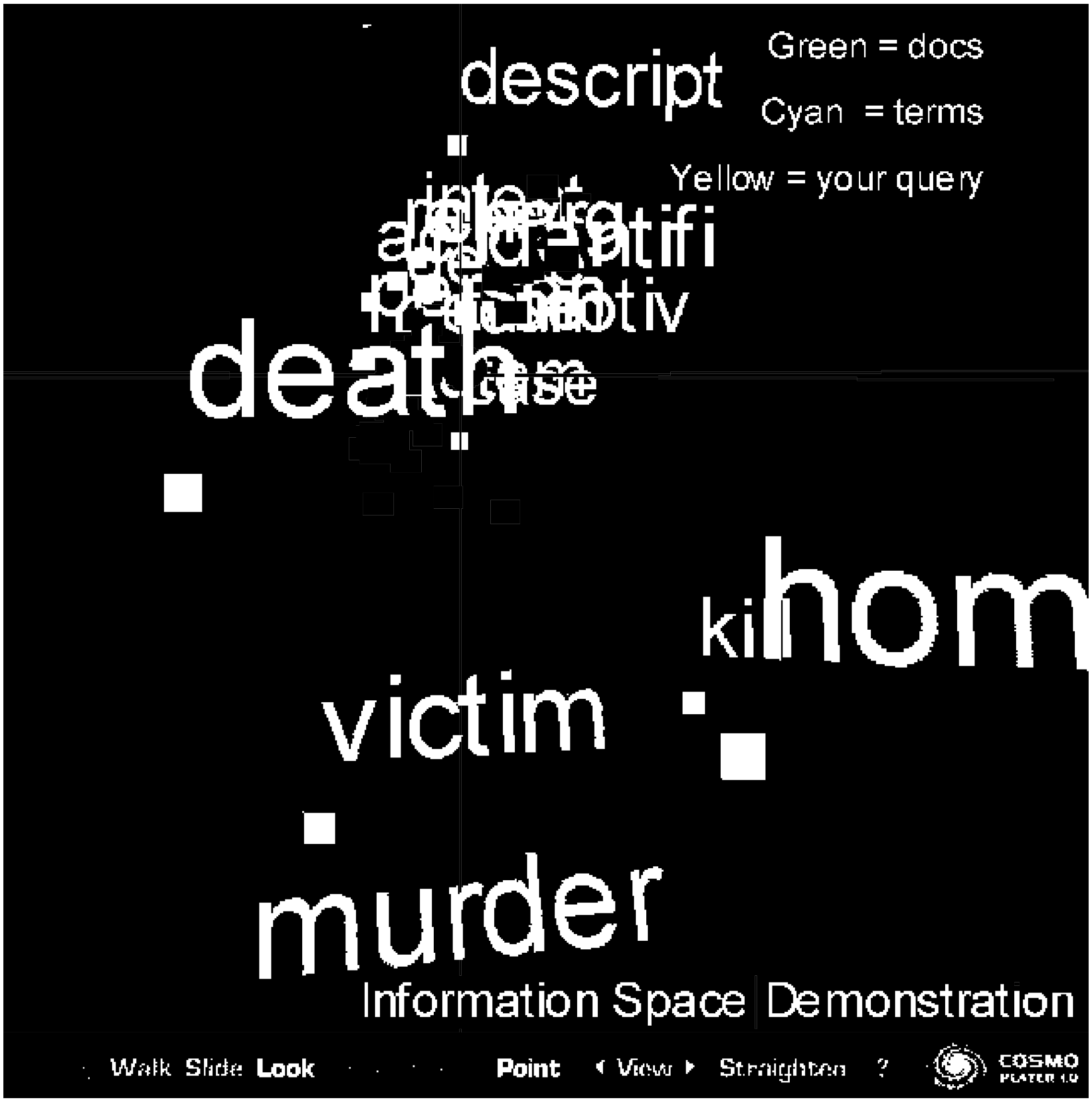
The second ranked additional set, based on F2, applied a 10% cutoff, but was otherwise comparable to the additional set based on F1. For the main F1 and F2 sets, as well as the additional ranked set for F1, the recall and precision scores were very low. Typically, an average of only 3% to 6% of the pooled relevant documents were retrieved per query. Exact precision was poor for the F1 sets (with maximum scores of .10), but better for the F2 sets (with a maximum of .44 for the unranked set, and .29 for the ranked set).

The filtering results are, in general, comparable to the routing results. The ability of the information space technique to identify non-relevant documents, in these experiments, was created more by the Boolean inclusion of documents based on query terms than on a particular cutoff distance from the query to the document.

The use of sub-documents did yield better recall and precision scores than were found for the routing task, which used entire documents. Further analysis will be completed to determine whether the retrieved sets of documents is appreciably different from sets retrieved by other TREC groups or by using full document information space methods. Numerically, the retrieved sets appear to be very different: less than 1/3 of the documents retrieved in routing were also retrieved in filtering, and vice-versa.

VISUALIZATION

Throughout the adhoc, routing and filtering tasks, retrieval was from a series of programs operating in batch mode. That is, each document collection, then every query, was run through a series of steps without human intervention. This section makes brief mention of one further application of the information space techniques described here: the visual interface.



The routing, adhoc and filtering task demonstrated capability of the context-based information space approach described here for clustering documents based on similarity, then retrieving documents based on their similarity to a query. Given that PCA extracts the largest eigenvalues first, it is possible to view the first three dimensions of the information space such that it is visually indicative of the overall multidimensional space. From 25% to 75% of the variance of the entire co-occurrence matrix is accounted for by these first three eigenvalues, enabling a reasonably accurate view.

The figure shows a VRML fly-through model of Query 189, accessed through the Cosmo Player plug-in to Netscape version 4. Query terms are visible; the query is towards the center of the space. The closest 100 documents to the query are displayed as dark squares, and may be clicked to retrieve the document.

The author has developed a Web-based interface to the entire information space-building system, with the capability of producing such a VRML world automatically. Because the process takes from 5 to 20 minutes, depending on the length of the query, this is not yet an alternative to existing real-time interfaces. The author is performing usability testing on the VRML interface, as well as an OpenGL version which functions with MSWindows and X-Windows.

DISCUSSION

These results for the adhoc, routing and filtering tasks show good improvement over the information space approach used in TREC-5. However, the results are far from outstanding. Further work is needed to develop term weighting techniques and methods for automatic query processing (expansion and truncation). The use of sub-documents seems promising, and would be especially interesting using manual relevance feedback or otherwise extracting relevant *passages* from previously judged relevant *documents*.

Some anomalies of the information space approach remain. The variability of the basic scale of the space (e.g., that the average inter-document distance varies greatly in different contexts) is clearly based partially on the number of query terms, but the nature of the relationship is not clear.

Visualization of information space is powerful for evaluating the clustering of documents, query terms and queries. Visual feedback systems are being developed that may be suitable for manual, versus automatic, participation in future experiments. Given the relatively unimpressive results for the three TREC-6 tasks described here, it may be that one of the main roles of the information space approach will be to create visual, spatial and navigable spaces based on response sets derived from traditional approaches. For example, a traditional Boolean, probabilistic or vector system might be used to generate a response set of potentially relevant documents for further processing and visualization using the information space approach.

Ongoing work includes further evaluation of the TREC-6 results to identify trends in the types of documents that are retrieved (length, presence of particular terms, etc.). In addition, queries with greater or lesser success will be examined to ascertain the relation between different types of queries and the effectiveness of the information space approach.

REFERENCES

Frakes, Willaim B. & Baeza-Yates, Ricardo, Eds. 1992. Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice-Hall.

Newby, Gregory B. (1997). "Metric multidimensional information space." In Harman, Donna (Ed.). Proceedings of the TREC Conference, volume 5. Gaithersburg, MD: NIST.