Metric Multidimensional Information Space

Gregory B. Newby[1]
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

The rationale and methodology for retrieval based on the relative locations of documents within a geometric information space are introduced. Results from category A routing and filtering experiments in TREC-5 are discussed. The techniques used are related to the vector space model, latent semantic indexing, and other methods that rely on statistical qualities of texts to assess document relatedness. Results show some promise, but additional research is needed to determine the extent to which retrieval may be improved over existing approaches.

INTRODUCTION

Spatial models for IR are well-known, with almost 30 years of refinement and variation. The primary points of departure of this work from many of the efforts utilizing the vector space model (VSM) or derivatives are:

> 1. Relations among terms are measured. This is contrary to te basic VSM, in which term vectors are mutually unrelated (orthogonal); and
>
> 2. Only a relatively small subset of available terms are utilized.

The first point is the most important; the second is more a matter of computational speed than of desire. The genesis of the specific methodology is based on a technique for the measurement of psychological phenomena through surveys called multidimensional scaling or MDS (KIM78, WF80). This technique results in a third departure from VSM:

> 3. The agreement between the information space and human cognitive space is subject to empirical verification.

This third quality will not be pursued in this paper, apart than to present the definitions of cognitive space and information space.

> *Information space* is a set of concepts and relations among them held by an information system. Information space is produced by a set of known procedures, and is changed through intentional manipulation of its content.

---

[1]GSLIS, UIUC; 501 East Daniel Street; Champaign; IL; 61820; USA. Tel: 217-244-7365; Email gbnewby@uiuc.edu.

> *Cognitive space* is the set of concepts and relations among them held by people or groups. Cognitive space is in constant change as a result of internal processes and interaction with the environment.

The nature of the cognitive approach to information systems is discussed extensively in INGW96. For this paper, it is sufficient to comment that the desire to develop methods for information space which approximate, in some ways, cognitive space is accepted as worthwhile. An example of an information system which would presumably benefit from such methods would be a customized information agent. For example, a computer program that selects items for presentation to a user from a constant stream of new documents based on *a priori* data on a particular user or group (such a program is indicated for the TREC-5 filtering task).

GENERAL METHOD

Retrieval from the information space as created here is based on the monotonic distance of documents from a query. A multidimensional information space is created based on statistical relations among terms; documents are then located at the centroid of their terms. Queries are also located at the centroid of their terms. Details on the steps involved follow.

Step 1: Select terms for the information space. The space will be built from a statistical analysis of term relations. Rather than assess relations among all unique terms in a set of data, it is desirable to identify a smaller number of terms to work with. For TREC-5, up to 2200 terms were used. Later improvements in programs allowed for up to 4000 terms. Further improvements are anticipated to yield capability for at least 10000 terms with current technologies. Throughout, all terms consist of only one word. It is anticipated that the use of multiple-word terms would yield interesting and useful results, but such an investigation has not yet been made. In addition, the statistical relations employed here could be applied to more complex "terms," such as the conceptual hierarchies described in SC95.

Selecting terms may be accomplished automatically or by consideration of some criteria. An automatic method might be to identify terms which are "frequent, but not too frequent." For example, between the 1st and 3rd quartile in the frequency distribution. (This presumes that a frequency distribution for the occurrence of terms has been calculated for a document collection or a representative subset.)

A non-automatic method for term identification might be to use terms from a collection that have been previously identified as being of interest. For example, a personal filtering system to seek network news (Usenet) articles of interest might use terms from documents that were judged to be of interest in the past. At this phase, a stoplist may be employed to de-select terms that are deemed of little value in discerning differences among documents. Stemming or truncation may also be applied.

For the TREC-5 work described here, a 500 word stoplist from SMART was used throughout. The only stemming used was to remove the letter 's' in the last word position. All words were truncated at 8 characters. It is anticipated that more intelligent stemming techniques could yield somewhat improved results (e.g., POR80).

Step 2: Generate a term by term co-occurrence matrix. The fundamental measure of "relatedness" applied in this method is the tendency of terms to co-occur in documents. Co-occurrence may be operationalized in a variety of ways for natural language documents. The outcome, regardless of the operational method, is an **N** by **N** matrix, where **N** is the number of terms selected in the previous step. The matrix is square and symmetric. Depending on the number of terms and the number of documents used to create the matrix, it may also be sparse (for TREC-5, no matrices were sparse).

The assumption underlying this method is simply that terms that tend to occur together are more closely related than terms that do not tend to co-occur. This is a much less sophisticated way of assessing term relations than is found in other settings. In expert systems, for example, measures of term relatedness include hierarchical relations, syntactic relations, temporal relations, and others (see, for example, PC88). Pathfinder networks use a more complicated measure of association than co-occurrence, yet without assigning semantic meanings to associations (MS99). Lelu (LELU91) applies a neural network.

Law & Whittaker (LW92) use the term 'co-word' method to refer to co-occurrence measures, as do PR93 and LEE93. The benefits of measuring simple co-occurrence is that it may be accomplished automatically for large document collections and that it is computationally simple and conceptually uncomplicated.

One way of operationalizing co-occurrence is to count the number of times each pair of terms in the matrix co-occur within a given (natural language) document. Thus, if **TERM(i)** and **TERM(j)** both occur in a document, the co-occurrence score for row **i** column **j** would be incremented, as would its symmetric pair, row **j** column **i**. This method essentially ignores the tendency for some terms to occur more than once in any given document, and also does not take into account the relative proximity of terms within documents.

A more detailed method would be to only increment a co-occurrence score for term pairs that occur within the same sentence, paragraph, document section, etc. Danowski (DAN88) has found that a sliding "window" of +/- 7 terms is effective in measuring term relatedness. Utilizing this method, terms that occur together in the same document would only contribute to the scores in a co-occurrence matrix when they occur closely together in that document. Terms which tend to occur more frequently in documents will have greater raw co-occurrence scores using this type of method.

A further operationalization of the creation of the co-occurrence matrix might be to post-process co-occurrence scores based on the relative frequencies of terms within documents and within the document collection. **tf * idf** is one common means of balancing the contributions of terms by examining their tendency to occur frequently within individual documents versus within a collection.

For TREC-5, the simplest approach was used: the co-occurrence matrix was built based on terms which occur in full-text documents, regardless of term or document frequency. Thus, a value of, say, 200 in the co-occurrence matrix for a given term pair would indicate that there are 200 document from the TREC-5 collection in which both terms occured.

Step 3: Extract a term space from the co-occurrence matrix. Principal components analysis (PCA) is applied to reduce the **N** by **N** matrix to a multidimensional term space. Principal components analysis is a multivariate statistical method similar to factor analysis (see KIM78 for a discussion). It is intended to identify trends in sets of data with large numbers of variables where none are pre-identified as dependent or independent. Eigenvalues and eigenvectors (which are used here as coordinates within a multidimensional space) are the primary outcomes of PCA.

Procedurally, PCA consists of iteratively identifying eigenvalues in the co-occurrence matrix. (More accurately, PCA first creates a correlation or covariance matrix from the co-occurrence matrix, then performs the eigenvalue extraction.) Pre-written computer functions and procedures for PCA are available for a variety of settings. The programs utilized in TREC-5 for PCA were written in FORTRAN and utilized the IMSL package. Since that time, the programs have been ported to C, but still utilize the IMSL routines **CORR** and **PRINC**.

The first eigenvalue and associated eigenvectors account for the most variance in the overall matrix of any possible linear combination. The second eigenvalue accounts for the most variance left over after the first, and so forth. By definition, a **N** by **N** matrix can be represented completely accurately by at most **N - 1** eigenvalues (for a real symmetric matrix). In most cases, however, there is enough co-variance in the matrix that fewer than **N - 1** eigenvalues are needed. As the extraction proceeds, the relative variance accounted for by each successive iteration decreases. It is typical, therefore, to only extract a limited number of eigenvalues by stopping with the cumulative variance accounted for reaches a desired threshold, such as 90 or 99%. PCA performed for TREC-5 typically accounted for 99% of the variance with 250-450 eivengectors for a 2000-term matrix.

Thus, a **N** by **N** matrix may result in a **D**-dimensional space, where **D** is far less than **N**. It is equally accurate to speak of the characteristics of the space in terms of eigenvalues and eigenvectors, matrices, or Euclidean geometry. The term space may be useful before documents are incorporated in the next step. For example, automatic query expansion or thesaurus lookups might be based on terms that are closest in the space to a target term.

Step 4: Locate documents in the space. Document locations in the space are calculated by placing them at the centroid of their associated terms. That is, a document is located at the geometric center of all of the **N** terms in the space that also occur in the document. For example, a document with 1000 unique terms might have 200 that are among the **N** terms in the space. The document would then be located at the center of the locations of those 200 terms.

This is the same approach used in the VSM. The difference is that the terms are not unit orthogonal vectors. A vector space with **N** terms would consist of **N** dimensions or vectors of equal size (typically, a size of one unit), each of which is mutually independent of the others. The difference in the current approach is that each term has a known quantified relation to every other term in the space. From a conceptual standpoint, this makes little sense because it is unreasonable to assert that terms are unrelated, but it eases computational approaches to assume this is the case.

Choices may be made about the specific method for locating the documents. For example, a document might be proportionally closer to terms that occur more frequently in that document by weighting that term more highly.

At the conclusion of this stage, the information space contains terms and documents. It is now ready for retrieval using various methods. It should be pointed out that empirical evidence through user studies (NEW93) has confirmed that one property of the information space is not what users would expect: although distance scores between terms and documents are known in the space, it is not the case that documents are located closely to their associated terms. This makes sense based on the steps taken above: by placing documents at the center of their associated terms, one would predict that a particular document would not be especially close to any of its terms. Users of a visual interface to such an information space, however, made the (reasonable) assumption that once a useful term in the space had been identified, all they needed to do was retrieve documents close to that term.

METHOD: TREC-5 DOCUMENTS

The information space created by the steps above is dependent on the terms selected for inclusion in the space, and by the documents used to create the co-occurrence matrix of those terms. The overall plan was to make a "context" information space for separate document collections. For example, a space could be built from the ZIFF collection (about 57K separate documents), and a separate space built from the Congressional Record collection (about 28K documents). Different sets of "interesting" terms could be derived from examining the frequency distribution of terms within each collection.

Each query (queries 251-300) could then be located in the separate information spaces and used to retrieve the closest documents to the queries in that space. In fact, this approach was used. One problem with the approach is that there is no good cross-space metric to decide how to mix documents from different collections. For example, in the ZIFF space the closest 100 documents might all be closer than .10 units, while in the Congressional Record space the closest 100 documents are all within .06 units. There's no sound basis, from the methods used here, to decide whether a document at, say, .05 units distance from the query in one space is equally "related" to a document at the same distance in another space. (Note that all the statistical method measures is "relatedness" based on the term co-occurrence matrices — there can be no claim that the distance measures "relevance" or anything else.)

Results of this method were not submitted, however. The author managed to confuse the requirements for the routing and ad hoc tasks, and ended up creating information spaces suitable for ad hoc queries but without realizing that the actual collection to be used was not derived from the "context" spaces. The result was that there was not enough time to complete the ad hoc tasks, nor to generate many separate information spaces. Instead, the process was as follows for both routing and filter tasks, which utilized only one information space:

Select first group terms for analysis. Unique terms from the routing and filtering queries[2] were determined (minus stoplist terms). 915 unique terms resulted.[3] All tagged fields from queries were dropped for all processing discussed here, and the entire query (*sans* tags) was used as-is with no further processing.

Select second group of terms for analysis. Results of the routing and filtering queries from prior TRECs were examined. 5727 of the documents that had been identified as relevant from prior TRECs were included in the collections on CD2 and 4 utilized for TREC-5. A frequency distribution of unique terms from these 5727 documents was created, and an additional 985 terms were selected from approximately the center of the distribution. These terms occurred in between 250 and 4000 of the 5727 documents. These 985 were added to the 915 previously identified, for a total of 1900 terms for further analysis. It is important to stress the non-scientific basis of this selection, as only term frequency was taken into account in order to select a manageable number of terms.

Build the co-occurrence matrix. Only the 5727 documents previously identified as relevant from prior TRECs were used to build the co-occurrence matrix. The idea was to create an information space based strongly on relevance. Again, this is an arbitrary decision: could it not be that the non-relevant documents would have been equally useful? Or, simply using the documents from other collections?

Perform Principal Components Analysis. PCA extracted 99% of the variance from the co-occurrence matrix, resulting in 454 eigenvalues (alternatively, a 454-dimension information space).

Locate documents. Each of the 130476 documents from the FB collection (the target for both the routing and filtering tasks) was located in the 454D space. All terms in each document that were also among the 1900 terms for analysis were identified. Documents were assigned coordinates at the (non-weighted) centroid of the terms.

Locate queries. The routing/filtering queries were located using the exact same technique as for documents, with the same pre-processing. Note that no query expansion techniques were utilized.

Select documents for "retrieval." Results for each query were generated. The geometric distance from each document to each query was measured. A ranked list of all 130476 documents for every query resulted.

Prepare result sets. The queries and collections were the same for the routing and filtering tasks.

---

[2]That is, queries with these numbers: 1, 3, 4, 5, 6, 11, 12, 23, 24, 44, 53, 54, 58, 68, 77, 78, 82, 94, 95, 100, 108, 111, 114, 118, 119, 121, 123, 125, 126, 142, 154, 161, 173, 185, 187, 189, 192, 193, 194, 195, 202, 207, 211, 221, 222, 224, 228, 237, 240 and 243.

[3]Throughout, documents and queries were pre-processed to remove all punctuation, change to lower-case only, remove trailing 's,' and keep only alphabetical characters (remove symbols and numbers). The impact of these hatchet techniques is difficult to assess without further experiment.

For routing, the closest 1000 documents for each query were submitted as results. For filtering, there were three sets. In the task guidelines, these were presented as: “Run 1 should be a high precision run, Run 3 a high recall run, and Run 2 somewhere in between.”

Since, as mentioned earlier, the information space only measures “relatedness,” a further arbitrary decision was made that Run 1 should be smaller, Run 2 larger, and Run 3 largest. Thus, the closest 25 documents were presented for Run 1, the closest 250 for Run 2, and the closest 500 for Run 3.

RESULTS

Performance of the information space techniques as presented here were fair, but far from outstanding. TREC-5 results are labeled “ISpace” (for lack of a better acronym).

Table 1: Ispace routing results compared to median relevant retrieved @ 1000

Below median	34
On median	6
Above median	5

Ispace seemed to do well at including relevant documents at the median or above when the total judged relevant was small (i.e., < 10). Overall, at least ½ of the total relevant documents were retrieved as part of the response set. Otherwise, there is no obvious pattern to the results.

Table 2: Summary of routing results across queries

Query number	Relevant	Relevant retrieved	Precision at 10 docs	Precision at 100 docs	Precision at 1000 docs	Exact Precision
1	30	15	.10	.010	.015	.033
3	101	26	.00	.01	.026	.0099
4	178	105	.10	.03	.105	.0281
5	19	13	.00	.00	.01	.00
6	158	49	.00	.02	.049	.0190
11	92	78	.00	.00	.078	.00
12	228	171	.00	.06	.17	.0658
23	7	4	.00	.00	.004	.00

24	38	16	.00	.00	.016	.00
44	10	6	.00	.00	.006	.00
53	1	1	.00	.00	.001	.00
54	48	38	.00	.01	.038	.028
58	45	12	.00	.00	.012	.00
77	14	7	.00	.01	.007	.00
78	37	21	.00	.01	.0210	.00
82	55	49	.00	.00	.049	.00
94	56	17	.00	.00	.017	.00
95	93	23	.00	.00	.023	.00
100	157	79	.00	.02	.079	.0127
108	174	101	.00	.12	.101	.0977
111	887	458	.60	.62	.458	.4307
114	42	14	.00	.01	.014	.0238
118	324	23	.0255	.00	.00	.0015
119	185	13	.00	.03	.013	.0216
123	57	22	.00	.02	.022	.0175
125	6	3	.00	.00	.003	.00
126	18	11	.00	.01	.011	.00
142	808	450	.50	.65	.45	.3837
154	22	7	.00	.00	.007	.00
161	153	26	.00	.03	.026	.0261
173	15	5	.00	.00	.005	.00
175	14	10	.00	.00	.01	.00
187	194	82	.00	.03	.08	.0206
189	584	264	.30	.37	.264	.1884
192	10	6	.00	.00	.006	.00
194	8	3	.00	.01	.003	.00

202	583	250	.50	.50	.25	.247
207	1	1	.00	.00	.001	.00
211	2	0	.00	.00	.00	.00
221	193	32	.00	.02	.032	.0311
222	2	2	.00	.01	.002	.00
224	1	0	.00	.00	.00	.00
228	68	43	.00	.02	.043	.0294
240	88	11	.00	.00	.011	.00
243	2	2	.00	.01	.002	.00

Precision ranged from extremely poor (exact values of 0 for 25 queries) to nearly acceptable (over .20 for 5 queries). Queries for which performance was acceptable include 111, 142, 189 and 202.

Table 3: Results over all 45 topics:

Query	Relevant	Relevant Retrieved	Precision at 10	Precision at 100	Precision at 1000	Exact Precision
all 45	1403	542	.05	.05	.03	.0299

Filtering

Filtering task results, at first glance, were not much different than the results for the routing task. Of the 7 groups that tackled this task, though, Ispace would seem to have performed the most poorly across the board. Of the 45 queries, Ispace achieved the lowest precision score on all three sets (precision oriented; balanced; recall oriented) for 20. That is, none of the other groups received a lower score for any of the those 20 queries.

For the remaining 25 queries, Ispace achieved the lowest score in at least one of the three sets for an additional 18 queries. Ispace had the highest score for a set for only 2 queries, and achieved a score close to the median on at least one set for an additional 4 queries.

An easy explanation for the poor performance of Ispace on many queries is forthcoming. In queries that resulted in very few relevant documents overall, the fixed number of documents retrieved resulted in very low scores for the three sets. 13 queries resulted in 10 or fewer documents judged as relevant. Each of these queries were among the 20 on which Ispace had the lowest achieved score.

Table 4: Summary filtering task results

	Precision oriented	Balanced	Recall-oriented
Mean precision	.2347	.1065	.0788
Mean recall	.0958	.3008	.3716

Analysis of the global statistics for all routing and filtering tasks indicates that other groups may have also had troubles due to the relatively small number of relevant documents found for the queries. Note that mean number of relevant documents per query is 129.1, but the standard deviation is 204.1. 15 queries had 20 or fewer relevant documents. This low number of relevant documents would result in low precision scores for systems which produced a ranked list of 1000 documents per query for the routing results, or who chose (as was the case for Ispace) to produce fixed-sized sets for the filtering tasks.

Table 5: Summary of routing and filtering task results for all of TREC-5 compared to Ispace

	Mean	Standard Deviation	Sum
Relevant	129.1	204.1	5808
Rel_Ret by Ispace	57.08	104.6	2569

Correlation Analysis

Analysis of Pearson product moment correlation coefficients across the set of scores produced for Ispace's routing results is slightly informative. Coefficients tend to be over .90 for almost all score pairs when all scores are compared (scores for relevant, rel_ret, recall at .00 - 1.0, precision at 5 - 1000 documents, average precision, and exact precision). This would be expected, since most scores rely on rel_ret. The only interesting feature of the correlation table (not presented here) is that correlations tend to drop off for recall after the .50 level, with many non-significant or small values. One interpretation of this is that the cluster of documents that Ispace places near the query is effective, but only to a point. Later-retrieved documents (documents further from the query location) are less likely to be relevant, yet other relevant documents *do* exist in the information space.

It may be inferred from this tendency that the measure of "relatedness" that Ispace is sensitive to is not entirely consistent with the measure of "relevance" as produced by the TREC-5 evaluators. Of course, that can be said of any system that does not achieve perfect results! For Ispace, an interesting question is whether the relevant documents not found close to the query location tend to cluster together in other locations in the information space.

Queries

An analysis of the queries themselves may be illuminating. Factors which might contribute to the effectiveness of Ispace include the length of queries, the presence of query terms in the documents of the information space, and the number of relevant documents from prior TRECs that were used to build the information space.

From the method as described herein, only one global information space was built for the routing and filtering tasks. The input to the space was the collection of documents that had been identified previously as relevant and were part of the CD2 or CD4 collections.

Table 6: Summary statistics for numbers of query terms and relevant documents used to build the information space

Figures for 45 TREC-5 routing/filter task queries	Number of terms in each routing query	Number of relevant document per query used for training from prior TRECs	Total relevant docs found for all TREC-5 routing/filter participants	Ispace total relevant docs for routing tasks
Mean	120.7	442.8	129.1	57
Median	116	372	48	16
Minimum	8	66	2	0
Maximum	522	1371	887	458
Standard Deviation	89	276.9	204.1	104.6

The figures in table 6 demonstrate variety in query sizes, and also that overall, Ispace seems to be capable of retrieving about ½ of the relevant documents. One item which is strongly supported is that the number of query terms is not related to the number of documents identified by TREC-5 evaluators as relevant ($p > .10$), nor is the number of query terms related to the number of relevant documents retrieved by Ispace ($p > .10$).

Information Space Input Documents

The interesting, but inconclusive, component of the correlation table of query length and number of documents retrieved and ranked as relevant is that there is a strong relationship between the number of documents that went into building the information space from each query and the number of relevant documents that Ispace eventually retrieved (r=.57, $p < .001$). Taken alone, one would infer that the context of the information space is extremely important — that including the relevant documents from past experience is very useful in improving results.

But the high correlation between the number of documents that went into building the information space and the number of relevant documents retrieved overall (r=.61, $p < .001$) indicates a different interpretation. Essentially, this correlation indicates that queries with larger numbers of relevant documents in past TRECs tended to have larger numbers of relevant documents in TREC-5.

Table 7: Correlation among values for TREC-5 and Ispace

Pearson's r, probability	Relevant docs prior TRECs	Relevant docs TREC-5	Relevant docs by Ispace	# post-processed query terms
Rel docs prior	1.0 0.000			
Rel docs TREC-5	0.614 0.001	1.0 0.000		
Rel docs Ispace	0.568 0.001	0.954 0.001	1.0 0.000	
# query terms	0.172 0.257	0.136 0.373	0.123 0.417	1.0 0.000

Examination of the queries that Ispace was able to perform best at helps to give sustenance to the first interpretation, that the context used to build Ispace's information space improves retrieval for queries more strongly related to or derived from that context. The queries on which Ispace did best for both filtering and routing tasks are 111, 142, 189, and 202.

Examination of these queries shows they are unremarkable in terms of the number of query terms (ranging from 8 to 189). Yet two of the queries (142 and 189) had the largest number of relevant documents from prior TRECs that went into the context for Ispace's information space. The other two queries (111 and 202) were also above the 3rd quartile.

The greatest sustenance is given to the second interpretation, however — that Ispace tends to retrieve larger numbers of relevant documents when all of TREC-5 identified larger numbers. This is the correlation of .95 ($p < .001$) between the number of relevant documents identified by TREC-5 evaluators and the number of relevant documents that Ispace retrieved.

Examining the number of relevant documents from prior TRECs that were actually on CD2 and CD4 (as opposed to those that were judged as relevant in prior TRECs but not part of the collection used for TREC-5) is not helpful, as the proportion of documents judged relevant to those on CD2 and CD4 is relatively constant. However, a somewhat weaker correlation is found between the number of relevant documents Ispace found and the number of documents actually used to build the information space (r=.47, $p < .001$). The weaker correlation (.47 versus .56 for all relevant documents from prior TRECs) is just as likely to be a byproduct of the larger standard deviation for the second score (277, versus 105 for documents actually used) as anything else.

CONCLUSION

Taken together, these results indicate there is some utility in the Ispace approach. However, it is yet to be seen whether the differences in the approach to document processing, document representation, query representation, and the retrieval process from Ispace offers any substantial improvements over existing approaches.

Shortcomings of creating a metric multidimensional information space are primarily practical: the large matrices and moderate computational complexity of the principal components analysis, coupled with the pre-processing of documents and then locating them in the information space, all result in a process that is not amenable to real-time production. By sampling from the documents to create the information space (rather than examining all), and by applying singular-value decomposition or another reduction technique to reduce the size of the matrix, some speed-up in processing would be expected.

Further research should address, minimally, the following areas:

1. What difference in performance would result from adjusting the selection of input documents to generate the information space as follows:

   - Choose a small set of known relevant documents only
   - Choose a large random set of documents
   - Manually pick terms for the space which are suspected to be useful for discriminating among relevant and non-relevant documents

2. What is the effect of the number of terms in the information space (or how they are selected)?

3. Examine how documents identified as "relevant" for TREC-5 tend to cluster in the information space. A random pattern would indicate that there are qualities of relevance that are entirely missed by Ispace, while the presence of well-defined clusters that are not at the query centroid would indicate that Ispace is sensitive, but the query location is not the only worthwhile basis for retrieval.

4. Create information spaces based on actual measurement of cognitive space (i.e., through measurement via MDS surveys), and determine:

   - The differences in the information spaces that result for different user groups or situations
   - Whether such information spaces could be approximated algorithmically

- Whether better retrieval results may be obtained with such information spaces

Many other areas are of interest, such as the role of multi-word terms, exploration in the effects of different stemming and truncation, or use of knowledge about different parts of documents (e.g., examining whether using only an abstract would help). One area which has not been explored in the current paper is the potential for experimentation with visual interfaces for IR.

The information spaces created here are different from vector spaces in that the term vectors are non-orthogonal. Further, the PCA procedure that creates the spaces extracts eigenvalues such that the largest are chosen first. Thus, it is typical for the first 3 eigenvalues to account for 15 or 20% of the variance in the information space. These first 3 dimensions, then (or, another set of early dimensions) are suitable for visualization and navigation with a 3D interface.

The author has created such interfaces using flat-screen and 3D technologies, and has found them to be usable for actual retrieval (NEW92). Such an interface is not well-suited to most of the TREC-5 tasks, but could be employed for the interactive task. Such efforts remain for a future time.

REFERENCES

[DAN88] Danowski, James A. 1988. "Organizational infographics and automated auditing: Using computers to unobtrusively gather as well as analyze communication." in Goldhaber, G.M. & Barnett, G.A. (Eds.). Handbook of Organizational Communication. Norwood, NJ: Ablex.

[ING96] Ingwersen, Peter. 1996. "Cognitive perspectives in information retrieval interaction: Elements of a cognitive IR theory." J. Documentation 52(1): 3-50.

[KIM78] Kim, Jae-On & Mueller, Charles W. 1978. Factor Analysis. Beverly Hills: Sage.

[LW92] Law, J. & Whittaker, J. 1992. "Mapping acidification research: A test of the co-word method." Scientometrics 23(3): 417-461.

[LEE93] Lee, Joon Ho; Kim, Myoung Ho; & Lee, Yoon Joon. 1993. "Information retrieval based on conceptual distance in is-a hierarchies." J. Documentation 49(2): 188-207.

[LELU91] Lelu, Alain & Claire, Francois. 1992. "Information retrieval based on a neural unsupervised extraction of thematic fuzzy clusters." Neuronimes.

[MS88] McDonald, James E. & Schvaneveldt, Roger W. 1988. "The application of user knowledge to interface design." in Guindon, Raymonds (Ed.). Cognitive Science and its Applications for Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum.

[NEW93] Newby, Gregory B. 1992. "An investigation of the role of navigation for informatin retrieval." Proc. Amer. Soc. for Information Science Annual Meeting 20-25. Medford, NJ: Learned

Information.

[PC88] Parsaye, * & Chignell, Mark. 1988. Expert Systems for Experts.

[PR93] Peters, H.P.F. & van Raan, A.F.J. 1993. "Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling." Research Policy 22: 23-45.

[POR80] Porter, M.F. 1980. An Algorithm for Suffix Stripping. Program 14(3): 130-137.

[SC95] Strzalkowski, Tomek & Carballo, Jose Perez. 1995. "Natural Language Information Retrieval: TREC-4 Report." In: TREC-4 Proceedings. Gaithersburg, MD: National Institute of Science and Technology.

[WF80] Woelfel, Joseph D. & Fink, Edward L. 1980. The Measurement of Communication Processes: Galileo Theory and Method. New York: Academic Press.