Information Space Gets Normal

Gregory B. Newby
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27573
gbnewby@ils.unc.edu

Experiments are presented based on unofficial results for TREC-7. Eigensystems analysis of a term co-occurrence matrix is compared to eigensystems analysis of a term correlation matrix. For each matrix type, the effect of term weighting and document length normalization is assessed. Recall-precision curves and other TREC statistics indicate that the use of the correlation matrix improves performance regardless of what term weighting or document length normalization is used.

Introduction

This paper presents unofficial results of an IR experiment conducted as part of TREC-7 participation. The system used, Ispace, was an adaptation of the author's prior work (Newby, 1997; Newby, 1996). There were three variables studied in the experiment:

1. Eigensystems of term a co-occurrence matrix versus eigensystems of a term correlation matrix
2. Term weighting versus no term weighting
3. Document length normalization versus no normalization

Title-only runs were also included. The overall retrieval approach is comparable to latent semantic indexing (LSI; see Deerwester et al., 1991). Rather than representing documents as a collection of the terms they contain, the approach of Ispace is to build an "information space" based on relations among terms in the collection, and then locate documents at the center of their terms. The effect is to base document relations (that is, their relative locations in the information space) on non-independent term relations. Information space is defined here as the contents and relations among them held by an information system (cf. Ingwersen, 1996).

The information space is derived from a matrix of pairwise term relations. This term by term matrix is subjected to an eigensystems analysis, which reduces the dimensionality of the matrix by identifying and collapsing linear trends. The resulting matrix may be trimmed at *k* dimensions to simplify computation and remove "noise" in the full matrix. Each term in the matrix may be thought of as a row (or vector) in the *k*-dimensional matrix of eigenvectors. Retrieval proceeds by locating queries in the information space and retrieving those documents closest to the query based on geometric (Euclidean) distance.

Conceptually, this process is one of representing a term by its relation to other terms. This is an important point of distinction from approaches which assume orthogonality among terms (such as Boolean systems and the basic vector space model, although term weighting explicitly addresses term relations).

The balance of this paper presents the methodology and results for assessing the value of term correlation versus co-occurrence in the context of term weighting, document length normalization, and title-only retrieval.

Methodology

The full TREC-7 collection was used with topics 351-400. Porter stemming was applied. Because a full information space cannot be built for the collection, due to the large number of terms (over 180,000 after stemming), a subset of terms was selected. Terms were selected by:

- Choosing only noun and adjective terms from the 50 topics, as identified by Eric Brill's part of speech tagger (Brill, 1994). After stemming, this yielded 535 unique terms.
- Building a full 180K by 180K term correlation matrix and selecting the most closely related terms for each of the topic terms identified.
- Deselecting stopwords (a version of the SMART stoplist was used)

- Keeping the most frequently occurring terms, regardless of part of speech. An additional 599 terms were added.

The result was 1134 terms for further consideration, which is a comfortable number for eigensystems analysis in a dense matrix. This number of terms is arbitrary, and was chosen to benefit from expanding on the terms in the topics without greatly overshadowing them. An 1134 by 1134 term co-occurrence matrix was extracted from the full 180K by 180K matrix. In the matrix, terms that occur in the same document are counted as having co-occurred (this is a relatively simple measure of term relatedness).

Two information spaces were built from this matrix. The first was simply an eigensystems analysis of the co-occurrence matrix. This matrix was about 30% sparse. The first 10 eigenvalues accounted for about 25% of the variance. 1134 dimensions were identified, although only the first 300 were used.

The second information space was from an eigensystems analysis of the Pearson product moment correlation scores from the co-occurrence matrix. This matrix was completely dense (no 0 cells). The first 10 eigenvalues accounted for about 25% of the variance. 1062 dimensions were identified (due to redundancy in the correlation matrix), although only the first 300 were used. 300 is an arbitrary cutoff value that has been used in past LSI research, but is not further investigated here.

All 535K TREC-7 documents were located in the information space at the center of whichever of the 1134 terms they contained, with 3 variations for each type of space:

1. A variation with no term weighting or document length normalization
2. A variation with term weighting only
3. A variation with term weighting and document length normalization

Each of topics 351-400 was then searched with Ispace to retrieve the closest documents in each space. Topics were subjected to the same condition as the space variation (weighting and/or normalization). In addition, a title-only version for each topic was run to assess the viability of the Ispace approach for shortened topics.

TF*IDF weighting was relatively simplistic, as taken from Frakes & Baeza-Yates (1992, pp. 372-375). Term TF weights assign a score to the importance of each term in a particular document. The formula used for TF weighting is:

$$tf_{ij} = [\ \log2\ (freq_{ij} + 1\ )]\ /\ [\ \log2\ (length_{j})]$$

where:

$freq_{ij}$ = the frequency of the i'th term in the j'th document
$length_{j}$ = the number of terms in the j'th document

This formula has the property of ranging from greater than 0 to 1 (provided all documents have a length of at least 2 terms). Term IDF weights assign a score to the importance of a term in a collection. The formula used for IDF weighting is:

$$IDF_{i} = [\ (\log2\ N)\ /\ n_{i}\ ] + 1$$

where:

N = the total # of documents in the collection
n_{i} = the number of occurrences for the i'th term

This formula has the property of always being greater than 1. For the TREC-7 collection utilized here, the range was from 1 to about 21.

Results

These results were not judged by TREC assessors, so we cannot know for certain that a higher proportion of judged documents in the retrieved sets for each query would not have an impact. However, the results seem clear and in the direction anticipated, based on post-hoc analysis of TREC-7 Qrels files from NIST.

Table 1: Summary outcomes of different conditions. "Weight" is whether TF*IDF weighting was applied. "Normal" is whether unit length vectors were used. "Title" is whether only the topic title field was used. "Rel_Ret%" is the percentage of total relevant documents retrieved across all topics. "AveP" is the average precision (non-interpolated) across all relevant documents. "P @ 20" is the precision at 20 documents. "Rank" is how well this condition performed, based on AveP.

Matrix type	Weight?	Normal?	Title only?	Rel_Ret %	AveP	P @ 20	Rank
Co-occur	N	N	N	8.4	0.0060	0.0250	11
Co-occur	N	N	Y	9.9	0.0142	0.0190	9
Co-occur	Y	N	N	7.3	0.0047	0.0190	12
Co-occur	Y	N	Y	10.1	0.0118	0.0190	10
Co-occur	Y	Y	N	18.4	0.0254	0.0800	7
Co-occur	Y	Y	Y	12.9	0.0187	0.0450	8
Correlate	N	N	N	21.1	0.0807	0.2010	3
Correlate	N	N	Y	31.1	0.0452	0.1020	6
Correlate	Y	N	N	29.1	0.0731	0.1780	4
Correlate	Y	N	Y	29.7	0.0455	0.1010	5
Correlate	Y	Y	N	46.5	0.1476	0.3110	1
Correlate	Y	Y	Y	38.7	0.0958	0.1760	2

Table 1 shows that all conditions in which the correlation matrix was used, even with no document length normalization or weighting, outperformed all conditions when only the co-occurrence matrix was used.

Within each matrix type, the best performance was achieved with document length normalization and term weighting. Title-only runs under-performed relative to their full-topic counterparts, which tends to support notions of LSI-like approaches being particularly well suited to "query by example," when a long query or existing relevant document is available.

Figure 1: Recall-Precision curves for the various categories. "pc" indicates Pearson Correlation, "ev" indicates eigensystems on co-occurrence matrix only. N=normalization, W=weighting, t=title-only.

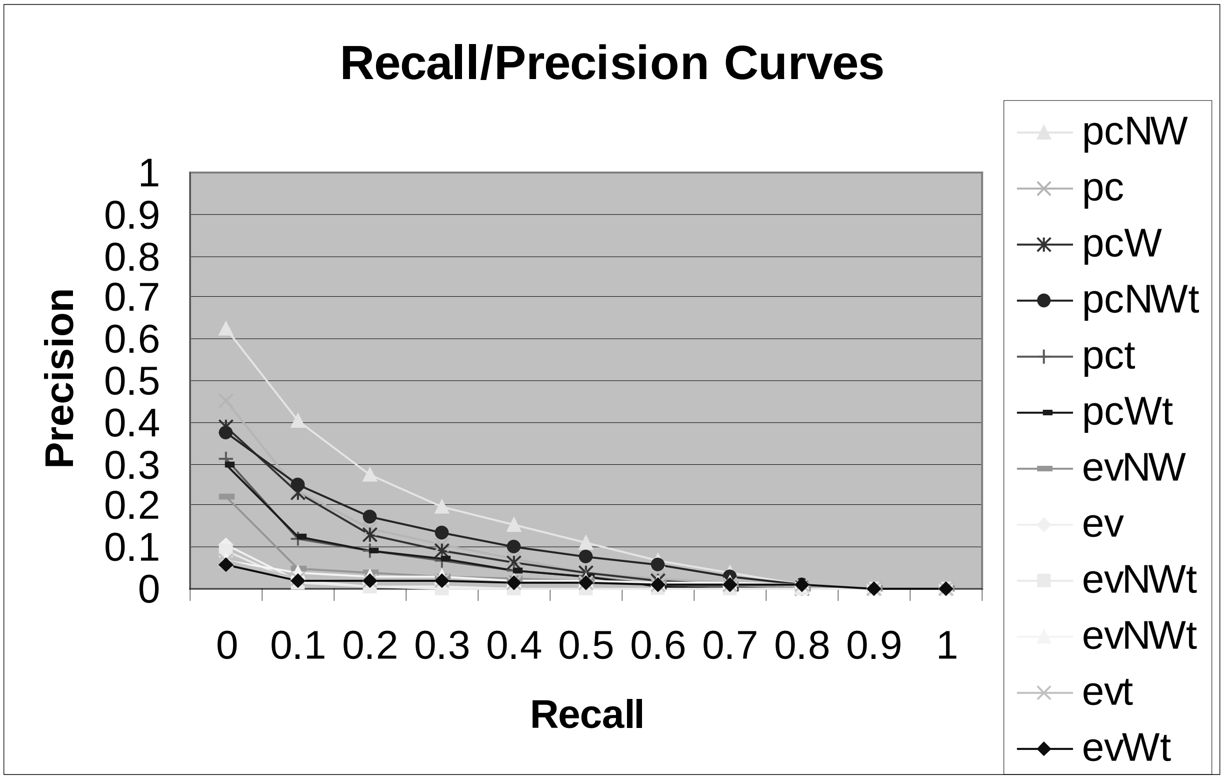

Figure 1 shows the relations among the various conditions graphically. Although the rankings are not quite the same as when only AveP is taken into account (Table 1), they are comparable and reinforce the value of term weights and document length normalization.

Conclusion

The benefits from document length normalization and term weighting that have proven themselves in many retrieval systems and have been shown to be effective here as well. Although this is confirmation of an expected result, not discovery of a new principle, the confirmation is a necessary step towards further progress. A reasonable expectation is that more advanced term weighting methods will further benefit retrieval effectiveness for the information space techniques discussed here.

A result that makes sense but had not previously been confirmed is the utility of computing the correlation matrix, rather than taking eigenvectors from the co-occurrence matrix alone. This approach, which is nearly identical to Principal Components Analysis (PCA) or Multidimensional Scaling (MDS, see Kruskal & Wish 1978), as employed in social science and elsewhere, pre-identifies linear term relations before the eigensystems analysis is performed. Although all the data required to build the correlation matrix is contained in the co-occurrence matrix, we add value by completing the correlation analysis.

Further research with Ispace will emphasize both practical and conceptual matters. Conceptually, modeling of the term relation process is in order. How is the LSI approach of examining a (sparse) term by document matrix different from the Ispace approach of analyzing a term by term matrix (the mathematical relation is well-known, but the conceptual relation less so)? What is the impact of choosing a far greater or smaller number of terms for inclusion in the information space? To what extent are non-orthogonal term vectors different than weighted but orthogonal term vectors as found in other IR approaches?

Practical matters need to focus on the efficiency of operation. Instead of only analyzing the documents that contain query terms, as is the case with almost all existing systems, Ispace requires evaluating each document in the collection relative to a query – essentially, a nearest neighbor search with hundreds of thousands of items in *k*-dimensional space. Other practical issues include the implementation of modern term weighting schemes, instead of the simple scheme used here, and empirically determining good cutoff values for the number of dimensions to keep. Different approaches for measuring term co-occurrence may also prove valuable.

References:

Frakes, William B. & Baeza-Yates, Ricardo (Eds.). 1992. Information Retrieval Data Structures & Algorithms. Englewood Cliffs, New Jersey: Prentice-Hall.

Brill, Erik. 1994. “Some advances in rule-based part of speech tagging.” Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Washington.

Deerwester, Scott; Dumais, Susan; Landauer, Thomas K.; Furnas, George W.; & Harshman, Richard A. (1990) “Indexing by latent semantic analysis.” Journal of the Society for Information Science 41(6), 391-407.

Ingwersen, Peter. 1996. “Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory.” Journal of Documentation 52 (1): 3-50.

Kruskal, Joseph B. & Wish, Myron. 1978. Multidimensional Scaling. Beverly Hills: Sage.

Newbym Gregory B. 1996. “Metric Multidimensional Information Space.” Text REtrieval Conference (TREC-5) Proceedings. Gaithersburg, MD: National Institute of Science and Technology.

Newby, Gregory B. 1997. “Context-Based Statistical Sub-Spaces.” Text REtrieval Conference (TREC-6) Proceedings. Gaithersburg, MD: National Institute of Science and Technology.