The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems

Helen J. Peat and Peter Willett*
Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Term cooccurrence data has been extensively used in document retrieval systems for the identification of indexing terms that are similar to those that have been specified in a user query: these similar terms can then be used to augment the original query statement. Despite the plausibility of this approach to query expansion, the retrieval effectiveness of the expanded queries is often no greater than, or even less than, the effectiveness of the unexpanded queries. This article demonstrates that the similar terms identified by cooccurrence data in a query expansion system tend to occur very frequently in the database that is being searched. Unfortunately, frequent terms tend to discriminate poorly between relevant and nonrelevant documents, and the general effect of query expansion is thus to add terms that do little or nothing to improve the discriminatory power of the original query.

Introduction

A major problem in the searching of natural language document databases is the need for the user to identify all of the terms that describe the subject of interest, so that it is possible for the search to differentiate correctly between the relevant and the nonrelevant documents for that query. A user's original query statement will typically consist of just a few terms germane to the topic and it is often necessary to add synonyms, variant spellings, etc. of the original set of search terms to achieve an effective search. This process is generally referred to as *query expansion* and has traditionally been carried out by means of thesauri and controlled vocabularies. The construction of these is extremely time-consuming and there has thus been considerable interest in techniques for the automatic identification of pairs, or groups, of words that are statistically associated with each other. The basic assumption underlying this work is that pairs of words that occur frequently together in documents are about the same subject; thus term cooccurrence data obtained from the analysis of a document collection can be used to identify some of the semantic relationships that exist between terms. More precisely, the Association Hypothesis states that "If an index term is good at discriminating relevant from nonrelevant documents then any closely associated index term is also likely to be good at this" (van Rijsbergen, 1979). It is generally assumed that terms used in queries are good at discriminating relevant from nonrelevant documents, so that closely associated terms, i.e., tems that cooccur frequently with the query terms, are also likely to be good discriminators and should thus be added to the original query. These additional terms may hence allow the retrieval of relevant documents that would not have been retrieved using the original query. This simple idea has now been studied for almost three decades.

Early experiments by Maron and Kuhns (1960) and by Stiles (1961) demonstrated the potential of term cooccurrence data for the identification of search term variants and this led to a large but rather unsystematic body of research that is summarized in the report by Stevens et al. (1965). It was not until the end of the Sixties that more detailed studies were carried out using standard document test collections. Lesk (1969) expanded a query by the inclusion of terms that had a similarity with a query term greater than some threshold value of the cosine coefficient (Salton & McGill, 1983). Lesk noted that query expansion led to the greatest improvement in performance when the original query gave reasonable retrieval results, whereas expansion was less effective when the original query had performed badly (a result that is, of course, in accord with the Association Hypothesis). An extended series of experiments was carried out by Sparck Jones (1971) on the 200-document subset of the Cranfield test collection. The terms in this collection were clustered using a range of different techniques and the resulting classifications were then used for query expansion. Sparck

*To whom all correspondence should be addressed.

Received February 6, 1990; revised April 6, 1990; accepted April 10, 1990.

Jones' results suggested that expansion could improve the effectiveness of best match searching if only the less frequent terms in the collection were clustered, with the frequent terms being unclustered, and if only very similar terms were clustered together. This improvement in performance was challenged by Minker et al. (1972), who worked with two test collections and with queries that had been expanded by between under 10% to over 200%.

More recent work on query expansion has been based on probabilistic models of the retrieval process and has tried to relax some of the strong assumptions of term statistical independence that normally need to be invoked if probabilistic retrieval models are to be used (Croft & Harper, 1979; Robertson & Sparck Jones, 1976). In a series of papers, van Rijsbergen and his coworkers have advocated the use of query expansion techniques based on a minimal spanning tree (MST), which contains the most important of the interterm similarities calculated using the term cooccurrence data and which is used for expansion by adding in those terms that are directly linked to query terms in the MST (Harper & van Rijsbergen, 1978; Smeaton & van Rijsbergen, 1983; van Rijsbergen, 1977; van Rijsbergen et al., 1981). Experiments with two large test collections showed that relevance feedback searches using expanded queries were noticeably superior to simple best match searches for which no relevance data were available (van Rijsbergen et al., 1981) but it was not clear whether this improvement was due to the expansion or to the feedback. Later work compared relevance feedback using both expanded and nonexpanded queries and using both MST and non-MST methods for query expansion on the Vaswani test collection (Smeaton, 1982; Smeaton & van Rijsbergen, 1983). It was not found possible to obtain consistent improvements in performance by the use of any of the query expansion methods; indeed, many of the expanded searches were noticeably inferior to the corresponding nonexpanded queries and it was also noted that adding randomly selected terms often gave better results than adding terms based on cooccurrence data. Smeaton and van Rijsbergen (1983) suggest that these disappointing results are due to the limited amount of data that is available when a relevance feedback search is to be carried out, a conclusion echoed by Yu et al. (1983).

The weight of the experimental evidence to date hence suggests that query expansion based on term cooccurrence data is unlikely to bring about substantial improvements in the performance of document retrieval systems. In this article we provide a rationale for this behavior that derives from the characteristics of the coefficients used to identify similar terms.

Frequency Characteristics of Terms and of their Nearest Neighbors

The studies referenced in the first section of this article have used a range of techniques to identify terms that are similar to query terms and that should be added to the query. However, they have all involved some way of calculating the degree of similarity between pairs of terms. This is usually a similarity coefficient, such as the cosine, Dice or Tanimoto coefficient (Salton and McGill, 1983): given two terms X and Y occurring in $F(X)$ and $F(Y)$ documents, respectively, these coefficients are defined to be

$$\text{COSINE}(X,Y) = \frac{F(X,Y)}{\sqrt{F(X) \times F(Y)}},$$

$$\text{DICE}(X,Y) = \frac{2 \times F(X,Y)}{F(X) + F(Y)}$$

and

$$\text{TANIMOTO}(X,Y) = \frac{F(X,Y)}{F(X) + F(Y) - F(X,Y)}$$

where $F(X,Y)$ is the number of documents in which X and Y cooccur.

It will be noticed that these coefficients are symmetric in $F(X)$ and $F(Y)$ and that their maximum possible values will be obtained when

$$F(X,Y) = \min\{F(X), F(Y)\}.$$

Given a specific term X, this means that the coefficients will have their maximum possible values for those terms, Y, which have frequencies of occurrence, $F(Y)$, that are the same as the frequency of occurrence for X, $F(X)$. For example, consider the cosine coefficient, which is probably the similarity coefficient that has been most extensively used in information retrieval research. The upperbound for this coefficient is given by

$$\text{COSINE}(X,Y) = \frac{\min\{F(X), F(Y)\}}{\sqrt{F(X) \times F(Y)}}$$

There are three possibilities

- $F(X) = F(Y)$.
  $\text{COSINE}(X,Y) = 1.0$;
- $F(X) < F(Y)$.
  Here, $\text{COSINE}(X,Y) = \sqrt{F(X)/F(Y)}$ and the value of the coefficient increases towards 1.0 as $F(Y)$ decreases towards $F(X)$;
- $F(X) > F(Y)$.
  Here, $\text{COSINE}(X,Y) = \sqrt{F(Y)/F(X)}$ and the value of the coefficient increases towards 1.0 as $F(Y)$ increases towards $F(X)$.

The fact that the largest values of the upperbound of $\text{COSINE}(X,Y)$ are obtained when $F(X)$ and $F(Y)$ are the same implies that the *nearest neighbor* for X, $NN(X)$, i.e., the term that is most similar to it, is likely to be one with a comparable frequency of occurrence, i.e.,

$$F(X) \approx F(NN(X)).$$

Hence, if X is a term in a query that is to be expanded, it is likely that the added terms will be of comparable frequency.

It is easy to demonstrate that this is indeed the case. We have taken the seven document test collections summarized in Table 1; these collections have been used in several previous research studies, both in our laboratory and elsewhere, and are discussed by Griffiths et al. (1986). For each of the terms, we have calculated the frequency and also the frequency of its nearest neighbor, using the cosine coefficient as the measure of interterm similarity and using the fast inverted file algorithm described by Willett (1981) for the calculation of these similarities. The product moment correlation coefficient between the sets of $F(X)$ and $F(NN(X))$ values was then calculated for each collection; these coefficients were as follows

- KEEN: 0.66
- CRANFIELD: 0.67
- EVANS: 0.67
- HARDING: 0.66
- LISA: 0.43
- SMART: 0.53
- UKCIS: 0.42

It will be seen that while there is a fair degree of variation in the precise value of the coefficient, it is also the case that there does seem to be a significant linear relationship between $F(X)$ and $F(NN(X))$ (in the ideal case of a perfect linear relationship, the value of the coefficient would be 1.00).

Discriminatory Abilities of Query Terms and of their Nearest Neighbors

The discussion in the previous section applies to all terms, whether or not they have been specified in a query. There is, however, one obvious difference between query terms and terms in general, viz. that the former tend to have substantially larger frequencies of occurrence than do the latter. This was first noted by Sparck Jones (1972) and forms part of the rationale for the well known inverse document frequency weighting scheme, in which the importance of a term is inversely proportional to its frequency of occurrence. Thus, Table 2 lists the mean frequencies for all of the terms and for just the query terms in the seven test collections of Table 1. Taken with the results of the previous section, the difference that is highlighted in Table 2 hence implies that the nearest neighbors of query terms will tend to have higher frequencies of occurrence than do the nearest neighbors of terms in general.

TABLE 2. The average frequencies of all terms and of query terms.

	Mean Number of Documents/Term	
Document Collection	All Terms	Query Terms
KEEN	5.8	17.3
CRANFIELD	15.9	61.6
EVANS	4.6	15.7
HARDING	10.3	55.6
LISA	17.9	236.6
SMART	25.2	361.4
UKCIS	9.1	115.7

So far, we have been considering only the frequencies of terms, without consideration of the extent to which these frequencies are related to the abilities of terms to discriminate between relevant and nonrelevant documents in a collection, i.e., their relative frequencies of occurrence in relevant and in nonrelevant documents as measured by their probabilistic relevance weights. The discriminatory abilities of terms has occasioned much research over many years (see, e.g., Salton, 1975; Sparck Jones, 1972; Willett, 1985): it appears that the relevance weights are greatest for low frequency terms, with the weights falling off inversely with term frequency (Biru et al., 1989; Croft and Harper, 1979; Robertson, 1986). The measure of discrimination used here is the Robertson and Sparck Jones relevance weight (Robertson & Sparck Jones, 1976), which has been used for this purpose previously by Biru et al. (1989) and by Smeaton and van Rijsbergen (1983). This weight is given by:

$$\ln\left(\frac{r_i(N - n_i - R + r_i)}{(R - r_i)(n_i - r_i)}\right)$$

where

- n_i is the number of documents containing term i
- r_i is the number of relevant documents containing term i
- R is the total number of relevant documents
- N is the total number of documents in the collection

The weight used the "0.5 modification" so as to avoid problems with zero-valued components (Robertson & Sparck Jones, 1976).

TABLE 1. Details of the seven document test collections that were used.

Test Collection	Documents	Terms	Queries	Terms/Document	Terms/Query
KEEN	800	1432	63	9.8	10.3
CRANFIELD	1400	2557	225	28.7	8.0
EVANS	2542	3730	39	6.6	27.5
HARDING	2472	8783	65	36.3	32.4
LISA	6004	13355	35	39.7	16.5
SMART	12684	18124	77	36.0	17.9
UKCIS	27361	20214	182	6.7	7.3

The discrimination weights were calculated for each term for each query in each test collection. These weights were then sorted into descending order to give a discrimination rank for all of the terms. The mean discrimination rank of the nearest neighbor to each query term was calculated for each document test collection, with the mean being calculated by averaging over all of the query terms for all of the queries. The mean discrimination rank was also calculated for terms picked at random from each of the collections; the number of randomly selected terms for each query corresponded to the mean number of terms in a query for that collection and thus the resulting mean rank was calculated using approximately the same number of terms as the mean rank for the nearest neighbors of query terms. The relationship between the frequencies of the nearest neighbors and their discrimination ranks are detailed in Tables 3–10. Tables 3–9 list the mean frequencies of query terms and the mean frequencies and discrimination ranks of their nearest neighbor for each of the seven collections. It will be seen that, in general, there is a close relationship between the two sets of frequencies and that the discrimination rank of the nearest neighbors increases in line with the frequency, i.e., the nearest neighbors become less and less discriminating. Thus, since most query terms are high frequency terms, as shown by Table 2, most of the nearest neighbor terms that would be added in by an expansion method would be poor, or very poor, discriminators that could be expected to do little to improve retrieval.

TABLE 3. Frequency and discriminatory characteristics for nearest neighbors of query terms in the KEEN collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	2.5	176.3
11–20	10.2	552.6
21–30	7.6	730.2
31–40	31.7	872.6
41–50	51.1	791.6
51–100	82.0	994.7
100–200	144.1	1139.7
200–500	96.2	1166.1

TABLE 4. Frequency and discriminatory characteristics for nearest neighbors of query terms in the CRANFIELD collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	2.2	253.5
11–20	10.3	736.6
21–30	24.7	1363.2
31–40	47.1	1299.3
41–50	47.5	1741.4
51–100	140.0	1976.8
100–200	279.3	1979.5
200–500	343.6	1932.5
501–1000	693.7	2324.2

TABLE 5. Frequency and discriminatory characteristics for nearest neighbors of query terms in the EVANS collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	1.8	452.3
11–20	2.6	737.5
21–30	8.2	1480.2
31–40	10.5	2063.6
41–50	7.2	2217.6
51–100	16.5	2667.9
100–200	14.9	2766.0
200–500	129.0	3289.0

TABLE 6. Frequency and discriminatory characteristics for nearest neighbors of query terms in the HARDING collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	2.4	1520.8
11–20	4.4	2776.8
21–30	6.0	4575.9
31–40	10.6	5141.2
41–50	19.7	5134.2
51–100	63.3	6469.0
100–200	301.8	7816.9
200–500	460.3	8093.0
501–1000	719.7	8333.4

That this is so is shown, in a rather dramatic fashion, by the figures in Table 10. Smeaton and van Rijsbergen (1983) noted that query expansion using randomly selected terms was often superior to (or, more accurately, not as bad as) query expansion using more rational expansion techniques; Table 10 demonstrates that this is because such terms tend to be noticeably better discriminators than the nearest neighbors, as measured by the former's lower discrimination ranks. The reason for this counter-intuitive finding is simple. The Zipfian distribution of term frequencies in document databases means that most terms have low frequencies of occurrence: randomly selected terms are likely to have low frequencies and discrimination ranks and thus to be better, i.e., not as poor as, discriminators than the high frequency terms that are likely to be identified by a conventional term expansion procedure.

The results listed in this article have been obtained using just the cosine coefficient to calculate the inter-term similarities and using just the nearest neighbors of the query terms. Peat (1989) presents additional, an-

TABLE 7. Frequency and discriminatory characteristics for nearest neighbors of query terms in the LISA collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	1.6	1268.4
11–20	5.7	2683.9
21–30	3.0	3987.8
31–40	8.4	6114.7
41–50	11.4	6245.8
51–100	143.8	8927.5
100–200	606.7	11098.8
200–500	1842.6	12576.9
501–1000	2767.1	12907.6
1000–5000	2427.1	12874.9

TABLE 8. Frequency and discriminatory characteristics for nearest neighbors of query terms in the SMART collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	2.1	1388.1
11–20	1.8	1942.8
21–30	21.2	6211.3
31–40	58.6	7233.8
41–50	28.7	6811.4
51–100	69.4	11185.0
100–200	390.5	13819.8
200–500	2238.5	16686.6
501–1000	2039.8	16828.2
1000–5000	3987.7	16936.4

alogous results that use the Dice and Tanimoto coefficients, that consider not just one, but also the five, 10, and 20 nearest neighbors for each of the query terms, and that use the expected mutual information measure (van Rijsbergen et al., 1981) to measure the discriminatory abilities of terms.

Conclusions

In this article, we have identified a substantial limitation in the use of term cooccurrence data as a basis for automatic query expansion in document retrieval systems. This limitation arises, in large part, from the characteristics of the coefficients that are used to measure the similarity between a pair of terms. In brief, we have shown that:

- a given term is likely to be most similar to terms that have comparable frequencies of occurrence in the document collection that is being studied;
- since query terms tend to have high collection frequencies, their nearest neighbors, i.e., those that are most similar to them and that are thus likely to be added to the query by an expansion method, are also likely to have high collection frequencies;
- since high frequency terms tend to be poor at discriminating between relevant and nonrelevant documents, the terms added to a query by an expansion method are unlikely to be effective discriminators

These findings thus provide a rationale for the lack of success that has been obtained in previous studies that have tried to use term cooccurrence data for query expansion. They can also be used to explain Sparck Jones' finding (1971) that the best retrieval results were obtained if only the less frequently occurring terms were clustered and if the more frequent terms were left unclustered and Smeaton and van Rijsbergen's finding (1983) that queries that had been expanded by the addition of randomly selected terms often gave better results than queries that had been expanded by any of their methods based on cooccurrence data.

We thus conclude (1) that the similarity coefficients currently available for automatic query expansion based on term cooccurrence data should be used only for the identification of alternatives to query terms that occur very infrequently in the database that is being searched and (2) the alternative terms that are identified by this procedure will also occur very infrequently.

Acknowledgments

We thank the Department of Education and Science for the award of an Information Science Advanced Course Studentship to HJP, the Computing Services

TABLE 9. Frequency and discriminatory characteristics for nearest neighbors of query terms in the UKCIS collection.

Frequency of Query Terms	Nearest Neighbors of Query Terms	
	Mean Frequency	Mean Discrimination Rank
1–10	5.2	4891.5
11–20	4.4	4493.6
21–30	3.8	3650.5
31–40	9.1	6847.1
41–50	48.9	11021.1
51–100	34.9	13323.8
100–200	169.6	16568.6
200–500	179.9	18360.0
501–1000	398.8	19408.3
1000–5000	348.3	19713.4

TABLE 10. Mean discrimination ranks for nearest neighbors of query terms and for randomly selected terms.

Collection	Nearest Neighbors of Query Terms	Randomly Selected Terms
KEEN	586.6	459.9
CRANFIELD	1650.3	1040.4
EVANS	1265.3	1271.1
HARDING	5653.0	3162.2
LISA	9946.9	4984.2
SMART	14174.9	6480.5
UKCIS	12958.6	6306.6

Department of the University of Sheffield for assistance with the extensive computations engendered by this study, and Janey Cringean, Alan Smeaton, and Tom Wilson for helpful comments on an earlier draft of this paper.

References

Biru, T., El-Hamdouchi, A., Rees, R., & Willett, P. (1989). Inclusion of relevance information in the term discrimination model. *Journal of Documentation, 45,* 85–109.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35,* 285–295.

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

Harper, D. J., & van Rijsbergen, C. J. (1978). Evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation, 34,* 189–216.

Lesk,.M. E. (1969). Word-word associations in document retrieval systems. *American Documentation, 20,* 27–38.

Maron, M. E., & Kuhns, J. K. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM, 7,* 216–244.

Minker, J., Wilson, G. A., & Zimmerman, B. H. (1972). Query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval, 8,* 329–348.

Peat, H. J. (1989). *An evaluation of the association hypothesis which underlies the idea of automatic query expansion in information retrieval systems,* University of Sheffield, MSc dissertation.

Robertson, S. E. (1986). On relevance weight estimation and query expansion. *Journal of Documentation, 42,* 182–188.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science, 27,* 129–146.

Salton, G. (1975). *A theory of indexing.* Philadelphia: Society for Industrial and Applied Mathematics.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval,* New York: McGraw-Hill.

Smeaton, A. F. (1982). *The retrieval effects of query expansion on a feedback document retrieval system,* University College Dublin, MSc thesis.

Smeaton, A. F., & van Rijsbergen, C. J. (1983). The retrieval effects of query expansion on a feedback document retrieval system. *Computer Journal, 26,* 239–246.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval,* London: Butterworth.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28,* 11–21.

Stevens, M. E., Giuliano, V. E., & Heilprin, L. B. (1965). *Statistical association methods for mechanized documentation.* Washington: National Bureau of Standards (Occasional Publication no. 269).

Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the ACM, 8,* 271–279.

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation, 33,* 106–119.

van Rijsbergen, C. J. (1979). *Information retrieval.* London: Butterworths.

van Rijsbergen, C. J., Harper, D. J., & Porter, M. F. (1981). The selection of good search terms. *Information Processing and Management, 17,* 77–91.

Willett, P. (1981). A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing and Management, 17,* 53–60.

Willett, P. (1985). An algorithm for the calculation of exact term discrimination values. *Information Processing and Management, 21,* 225–232.

Yu, C.T., Buckley, C. and Salton, G. (1983). A generalized term dependency model in information retrieval. *Information Technology: Research and Development, 2,* 129–154.