CHAPTER 2

The selection of good search terms

Introduction

This paper tackles the problem of how one might select further search terms, using relevance feedback, given the search terms in the query. The approach taken is based on an earlier paper by one of the authors [1] in which a theoretical model for the exploitation of statistical dependence between index terms was described. This model was evaluated in a later paper [2] showing the extent to which the use of statistical dependence information derived from co-occurring index terms would lead to an improvement in retrieval effectiveness. The experimental methodology used in the latter paper will also be adopted here, that is, the methods of evaluation and estimation will be the same except for minor variations which will be pointed out when appropriate.

The general experimental set-up within which the ideas for the selection of search terms were worked out can be simply described. There are a number of test collections consisting of documents, and queries with associated relevance assessments. For each query the relevant documents are therefore known, so that a user's response to any output from a retrieval strategy may be simulated. (Sometimes the relevance assessments are not exhaustive. This happens when only a portion of the entire collection assumed to contain most of the relevant documents is scanned in determining the documents relevant to a query. Unassessed documents are then assumed to be non-relevant.) The basic relevance feedback strategy is one in which a simple strategy such as co-ordination level matching is used to retrieve an initial small set of (say 10 or 20) documents. The known relevant documents in this small set are then used to estimate certain parameters, which in turn are used to build up a new search function. This new search function will incorporate new search terms, not already occurring in the query, which are derived from a tree relating all the index terms in the entire collection. The tree structure is in the nature of a thesaurus although the links are statistically derived. It is this tree structure called the maximum spanning tree (or MST briefly) - see appendix - which is the main aid used for finding further search terms. The way in which the spanning tree is actually used during a retrieval run is not very different from earlier uses of term clustering to expand queries. The relationship between an MST and a commonly used clustering method, single-link, is explained in [4]. Thus it is not too difficult to interpret the MST for index terms as a term clustering. But it should be stressed that the tree structure does not imply a hierarchical relationship between the terms. In fact each of the connected terms are informative about each other.

The MST lies at the centre of the research described in this paper. It is this structure which captures the important dependencies between index terms. In the original model [1] for the use of statistical dependence between index terms, two such trees were envisaged, one for the terms in the relevant documents and one for the terms in the non-relevant documents. Subsequently this was approximated by assuming the same tree structure for both sets. In this paper we go one step further and, for the time being, abandon the attempt to construct probability distributions on both the relevant and non-relevant sets, until we can resolve the difficulties inherent in making an explicit assumption on both relevant and non-relevant sets of documents. Instead we simply try to expand queries by appropriate index terms in the MST, which itself is based on the distribution of co-occurrences in the entire collection. Preliminary experiments with this way of expanding queries were reported in our earlier paper [2]. To avoid confusion between the way the MST has been used in earlier work and the way it is used here we shall now briefly discuss the relationship between the MST and the underlying probability model.

When faced with the problem of modelling the probability of relevance through distributional information about individual index terms one can make various assumptions about the independence or dependence of the index terms. A common one has been to assume that the index terms are independent on both the relevant and non-relevant sets of documents [3]. In [1] a limited form of dependence on both these sets was assumed. To capture the important dependencies a spanning tree is constructed for each set. However, attempts to use this form of dependence model run into estimation and computational problems which remain to be solved. Instead we have sought a compromise solution, one which would enable us to use the dependence information and yet not have to estimate this from ridiculously small samples. In this paper we attempt to make use of the statistical dependence between index terms over the entire collection. Assuming dependence on the entire collection is consistent with assuming independence on both the relevant and non-relevant sets [4]. In [4] it was shown how this particular form of conditional independence can lead to sensible heuristics for expanding queries by index terms connected into an MST based on co-occurrence data derived from the entire collection.

Our main concern in this paper is with the use of the MST connecting all index terms derived from distributional information about index terms in the entire collection. We use the spanning tree to expand the initial query. The effectiveness of this expansion is compared with that of the unexpanded query. The question naturally arises as to what is the 'best' spanning tree to use in this process. In [1] it was shown how the spanning tree might be constructed in an optimal way so that it produced the best approximation for the relevant probability functions. In that paper it was also conjectured that reasonable approximations to the optimal tree, although suboptimal, may yet give comparable performance in terms of retrieval effectiveness. That this is in fact so is demonstrated for several test collections.

The experimental results about the effectiveness of relevance feedback based on differently generated spanning trees are presented in the sequel. The feedback strategy evaluated is relatively simple, but constitutes a starting point for further research into more elaborate ways of using the spanning tree. The essential idea is to use a term contained in the query to lead us to further search terms. One can formulate the idea in terms of the Association Hypothesis [4]:

> If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is also likely to be good at this.

When used without further qualification in expanding queries, one must obviously make the implicit assumption that query terms are good at discriminating relevant from non-relevant documents. Now although this is not always the case it is not unreasonable to assume it on the average. Thus we expand each query term by the terms that are closely associated with it, and those are to be found by looking at the adjacent terms in the MST, those which are connected to a query term. The union of the query and the sets of adjacent terms, one set for each query term, we called the expanded query in our previous paper [2].

Expanding the query terms by all the adjacent terms in the spanning tree is clearly not the only way of proceeding. One would like to be able to select in order of preference further search terms. This general theoretical problem of how one might choose search terms in order of decreasing discrimination power remains to be investigated.

Before we can discuss our experimental results we must briefly describe the theoretical framework which has led to this work. Much of it may be found in Chapter 6 of Van Rijsbergen [4].

Basic Symbols

The theoretical background to this work in IR derives from a straightforward application of Probability Theory including some simple use of statistical decision theory. To explain it we shall need to define a few symbols.

$$\underline{x} = (x_1, \dots x_n)$$

represents a document, where n is the number of index terms in the vocabulary and $x_i=0$ when the ith term is absent and $x_i=1$ when the ith term is present. We consider only two relevance categories,

w_1 : relevant
w_2 : non-relevant

The important probabilities we need to define are,

$P(\underline{x})$

is the probability of observing a document description $\underline{x}$ within the document collection irrespective of whether it is relevant or not.

$P(\underline{x}|w_i)$

is the probability of observing $\underline{x}$ given that it represents a relevant ($i=1$) or non-relevant ($i=2$) document.

$P(w_i|\underline{x})$

is the probability that a document is relevant ($i=1$) or non-relevant ($i=2$) given its description $\underline{x}$.

$P(w_i)$

is the prior probability of observing a relevant ($i=1$) or non-relevant ($i=2$) document.

Obviously

$$P(\underline{x}) = P(\underline{x}|w_1)P(w_1) + P(\underline{x}|w_2)P(w_2)$$

Optimality

The fundamental assumption made in all this work is that the distribution of descriptions on the relevant documents is different from the distribution of descriptions on the non-relevant documents <u>and</u> that the difference can be estimated and used to find relevant documents. The main quantity estimated for finding the relevant documents is $P(w_1|\underline{x})$ i.e. the probability of relevance for every document. The higher the probability the more likely we are to want to retrieve that document. (From now on documents will be identified with their descriptions unless the difference is important.) The simplest retrieval rule using these probabilities is given by the following,

$$P(w_1|\underline{x}) > P(w_2|\underline{x}) \;\to\; \underline{x} \text{ is relevant, } \underline{x} \text{ is non-relevant} \qquad \text{D1}$$

This is a good rule for the following reason: it minimises the <u>expected</u> probability of misclassification (sometimes called the error rate). The probability of misclassification is given by

$$\begin{aligned} P(\text{error}|\underline{x}) = {} & P(w_1|\underline{x}) \text{ if we decide } w_2 \\ & P(w_2|\underline{x}) \text{ if we decide } w_1 \end{aligned}$$

So if for every document $\underline{x}$ we choose that w_i corresponding to the larger of $P(w_1|\underline{x})$ and $P(w_2|\underline{x})$ then the choice will minimise $P(\text{error}|\underline{x})$ for each $\underline{x}$. In doing so we will also minimise

$$P(\text{error}) = \sum_{\underline{x}} P(\text{error} \mid \underline{x})P(\underline{x})$$

which is the expected probability of misclassification.

In [4] it is shown that this approach can be generalised to incorporate different costs associated with different errors. That is, we can associate a different cost with retrieving a non-relevant document from missing a relevant document. The retrieval rule will then be expressed in terms of expected cost, and will specify the choice leading to minimum expected cost. However, this generality is not required here: interested readers should consult [4].

In practice the retrieval rule D1 is evaluated using Bayes' Theorem:

$$P(w_i \mid \underline{x}) = \frac{P(\underline{x} \mid w_i)P(w_i)}{P(\underline{x})} \qquad \text{Th1}$$

To evaluate D1, $P(w_i \mid \underline{x})$ is replaced by the R.H.S. of the equality in Th1. Furthermore, rather than estimate $P(w_i)$ when evaluating the arguments of D1, we prefer to rank the documents by $P(w_1 \mid \underline{x})$ thus obviating the need to set a cut-off. One can say more about the ranking than one would suspect at first sight. In fact there is an optimality principal now commonly known as the Probability Ranking Principle which states that ranking by $P(w_1 \mid \underline{x})$ is optimal in the sense that at any fixed recall level the precision will be maximised [5]. A simple proof of this can be found in Harter [6].

Independence

Let us look a little more closely at the retrieval rules that will result. Either way, whether we rank or use D1, we must estimate $P(w_1 \mid \underline{x})$ by the R.H.S. of Th1. The usual assumption made is that the index terms are statistically independent. To say this just like that is actually ambiguous without precisely specifying the sets on which the independence holds. So let us assume independence on both the relevant and non-relevant documents. Then we may write

$$P(\underline{x}|w_1) = P(x_1|w_1) \; \ldots \; P(x_n|w_1)$$
$$P(\underline{x}|w_2) = P(x_1|w_2) \; \ldots \; P(x_n|w_2)$$

Using these expressions we can derive the usual weighting functions [1] [2] [3], but let us first note that ranking w.r.t. $P(w_1|\underline{x})$ is the same as ranking w.r.t.

$$\log \frac{P(\underline{x}|w_1)P(w_1)}{P(\underline{x}|w_2)P(w_2)}$$

This last function is generally used to write down explicitly the weighting function as follows. Define

$$p_i = P(x_i=1|w_1)$$
$$q_i = P(x_i=1|w_2)$$

then

$$P(\underline{x}|w_1) = \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i}$$

$$P(\underline{x}|w_2) = \prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{1-x_i}$$

Now substituting in the log function above we get

$$g(\underline{x}) = \log \prod_{i=1}^{n} \frac{p_i^{x_i}(1-q_i)^{1-x_i}}{q_i^{x_i}(1-p_i)^{1-x_i}} + \log \frac{P(w_1)}{P(w_2)}$$

$$= \sum_{i=1}^{n} c_i x_i + \text{Const}$$

where

$$c_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

This $g(\underline{x})$ is the linear weighting function and c_i are the weights that are estimated from the relevance information contained in a small set of documents retrieved by a simple strategy such as co-ordination level matching. For example if the set of documents retrieved is N and we display the relevance information in the usual contingency table,

	Relevant	Non-relevant	
$x_i=1$	r_i	n_i-r_i	n_i
$x_i=0$	$R-r_i$	$N-n_i-R+r_i$	$N-n_i$
	R	$N-R$	N

we can derive the form of c_i used in [3], by using maximum likelihood estimates for p_i and q_i (i.e. $p_i=r_i/R$ and $q_i=(n_i-r_i)/(N-R)$) we get,

$$c_i = \log \frac{r_i/(R-r_i)}{(n_i-r_i)/(N-n_i-R+r_i)}$$

Although this form of c_i has proved successful, it causes difficulties when one or more of the interior cell values of the contingency table goes to zero, for then the log function is undefined. Robertson and Sparck Jones [3] sought to get around it by a well known statistical technique of adding .5's to the interior cells and adjusting the marginals accordingly. Unfortunately it does not solve the problem, for it grossly overestimates the probabilities involved.

Of particular interest is the case when a term is not assigned to any relevant documents ($r_i=0$). In the table below values for $100c_i$ are tabulated for a typical document collection: set N=1400, R=2, and the n_i and r_i values as shown in the table; adding .5's to the interior cells and adjusting the marginals of the contingency table will result in the table as shown.

	$100c_i$		
n_i	$r_i=0$	$r_i=1$	$r_i=2$
25	238	403	
50	168	331	568
75	125	288	494
100	95	257	450
125	71	233	419
150	51	212	394

In the absence of any confirming information of a term's ability to descriminate relevant from non-relevant documents, some large weights are computed. Compare the entries for ($r_i=0$, $n_i=25$) and ($r_i=1$, $n_i=125$). The reason for the large weight when $r_i=0$ can be traced to overestimating the parameter p_i. The '.5 technique' is equivalent to estimating p_i with $(r_i+.5)/(R+1)$. To illustrate the extent of this overestimation we now also tabulate some p_i values computed in this way when $r_i=0$ and R ranges from 1 to 5.

R	p_i	n_t
1	.33	467
2	.20	280
3	.14	200
4	.11	155
5	.09	127

The final column headed n_t shows values below which any value of n_i will lead to a positive value for the corresponding weight c_i. These tables were in fact tabulated for the Cranfield 1400 collection for which the average n_i per query term is only 169. Therefore, frequently (sometimes large) positive values of c_i will be computed when a term is not assigned to any relevant documents. Hence the '.5 technique' must be considered inadequate. For this reason we feel justified in proposing a different form of c_i.

In our previous paper [2] we suggested and evaluated a different form of c_i which we now realise will get around some of the estimation problems. The weight we suggested in [2] was

$$E_{iq} = \sum_{x_i, w_q} d_{iq} \, P(x_i, w_q) \log \frac{P(x_i, w_q)}{P(x_i) P(w_q)}$$

(The subscript q has been introduced to emphasise the fact that our calculations are always with respect to one query q.)

where

$x_i=0$ for absence and $w_q=w_1$ for relevant
$x_i=1$ for presence and $w_q=w_2$ for non-relevant

and d_{iq} is given conveniently by the following table:

	w_1	w_2
$x_i=1$	+1	-1
$x_i=0$	-1	+1

Rewriting E_{iq} as

$$G_{iq} = \sum_{x_i, w_q} d_{iq} D_{iq} P_{iq}$$

then $D_{iq}=P(x_i, w_q)$ is the 'degree of involvement' of cell (x_i, w_q) and P_{iq} equal to the log part is called the 'probabilistic contribution'.

It is easy to show that with $D_{iq}=1$ the weight G_{iq} simplifies back to c_i. But with D_{iq} equal to the joint probability one does not need to 'adjust' the cell values of the contingency table by .5's, since whenever an interior cell is zero we simply set 0log0=0 - which makes mathematical sense. We have no theoretical justification for this weight, but contrary to expectation, it outperforms the so-called independence weight c_i. Obviously, since c_i is optimal under the independence assumptions, this superiority must be related to the difficulties associated with estimating c_i from small samples. It may well be that, in the light of G_{iq}'s robustness and effectiveness, some theoretical justification will be found. It is this weight G_{iq} which is used in the experiments reported in this paper.

Dependence and the role of the MST

In deriving the linear weighting function $g(\underline{x})$ we made the assumptions of statistical independence on both w_1 and w_2. We could go on to say that these assumptions are unrealistic and therefore we would need to estimate the dependence between index terms to improve retrieval effectiveness. This is in fact what we set out to do in [1] and [2]. As we pointed out earlier, certain technical problems, mainly to do with estimation, prevented us from developing this approach further. However, a heuristic approach to using the maximum spanning tree has proved an interesting alternative to the strict term dependence model. For this we assume only that the terms are statistically dependent on the entire collection, and use the dependence to lead us from the query terms to further search terms.

The maximum spanning tree capturing the important dependencies is generated in a similar way to that described in [1] [2] [4]. Choosing an association measure from Table 1 we represent the pairwise association between index terms as a graph: the links measuring the association between any two terms represented by nodes. From this graph we can derive a maximum spanning tree, that is, a tree which spans the nodes in such a way as to have a maximum sum of links. So for each measure of association we will arrive at a different MST. Of these MST's one based on the expected mutual information measure (EMIM) is special. It is the basis for estimating in an optimal way the probability functions involved in our retrieval rule. The estimate is optimal in the sense that if we condition our variables in the way shown by the MST based on EMIM then we find a closer approximation to the underlying probability function than if we used any of the differently derived MST's. Although this result is not of immediate concern, it provides the motivation for (a) selecting an MST in the first place as a useful object, and (b) preferring certain MST's over others. In [1] it was conjectured that despite this optimality result, it may well be that a different, suboptimal but more efficiently generated, MST could give comparable retrieval performance to one generated from EMIM. In this paper we show this to be so.

The optimality of the MST based on EMIM deserves some further comment. We are assuming that if one models the underlying probability functions as closely as possible then one will get the best possible retrieval. The Probability Ranking Principle guarantees this for the strict dependence model [1]; its optimality is a matter of statistical fit. As soon as one breaks away from the strict dependence model, the role of the MST changes and the optimality of the EMIM-based MST is no longer guaranteed. One can only conjecture whether or not the MST based on EMIM is still the best possible.

A different approach

In this section we shall be more specific about the strategy used to expand queries with the aid of the MST and about its evaluation within a relevance feedback context. The general form of the weighting function is linear in the same way as the one derived from the independence assumption, but it is different in that the weights c_i are replaced by G_{iq}. The basic feedback strategy is to retrieve a small set of documents, 10 or 20, by choosing the 10 or 20 documents best co-ordinated with the query. This set is then used to estimate the G_{iq} weights for the query terms and the adjacent terms in the MST. The estimates required are for probabilities conditioned on either the relevant or non-relevant documents. In [2] we showed that the best way of doing this is to estimate $P(.|w_1)$ from the relevant documents in the feedback set and $P(.|w_2)$ from the entire set of documents minus the relevant documents in the feedback set. This is particularly important when estimating the probabilities for the probabilistic contribution in G_{iq}. The degree of involvement may be estimated in the same way although restricting the estimates to just the feedback set appears to be satisfactory [2]. Notice that whereas in [2] we felt obliged to adjust all our estimates by .5's here we have omitted to do this since we now realise that it is unsatisfactory and unnecessary. This will lead to minor discrepancies between precision and recall figures in this paper and corresponding ones in [2].

Three test collections, Cranfield 1400, UKCIS I and UKCIS II are used to measure the retrieval effectiveness of the feedback strategy under different conditions. The details of these test collections are summarised in Table 2. The method of evaluation, in terms of precision and recall, is the same as that in the earlier paper [2]. It is necessary to remind the reader that in evaluating feedback strategies we have adopted a method of residual ranking. Briefly, the feedback documents (seen by the user) are removed from the collection and precision recall figures calculated for the search on the remaining documents.

Experimental results

Our benchmark for the experiments is COORD(N) where N can be either 10 or 20, indicating the size of the feedback set. (The mnemonic 'name'(N) is used only in the text. In the tables columns will be headed by 'name', and the value of N will be given at the start of the table as (cutoff=N).) COORD(N) simply continues the co-ordination level match on the remainder of the document after the feedback documents have been removed. So there is no expansion and no feedback. All strategies employing feedback are shown to be superior to this benchmark.

Our first minor result is to establish the adequacy of the G_{iq} weight. For this we compare its performance on all three test collections with the independence weight c_i under the same condition: using both feedback and expansion. The expansion is done using the standard MST generated using EMIM. To see the difference the reader should consult Table 3 and compare the precision recall figures for EMIM(N) with those for IND(N). This clearly shows the superiority of using the G_{iq} weights on all three test collections. The result is the more remarkable for the fact that if the terms in the query and those added through expansion are assumed to be independent on both the relevant and non-relevant sets then theoretically the reverse should be the case i.e. IND(N) should be superior to EMIM(N).

The major result is a comparison of the feedback strategies, using linear weighting, expansion and the G_{iq} weight, where the expansion is done with MST's derived from the associated measures listed in Table 1. With each MST is associated a mnemonic identifying the appropriate association measure. This mnemonic is also used to identify the precision recall figures associated with the corresponding feedback experiment. These figures have been collected in Tables 4, 5 and 6, one for each test collection. For example Table 5 shows a column headed Maron in the first half of the table, and these figures pertain to a feedback strategy on UKCIS I with 10 documents in the feedback set and the MST based on the association measure labelled Maron in Table 1. No attempt has been made to graph the precision recall figures since we are only attempting to establish a no-difference effect. The figures bear out this claim: spanning trees generated from reasonable association measures do not give appreciably different retrieval results. It is interesting to note however that the MST derived from the EMIM measure on the whole gives slightly better retrieval effectiveness than all others, with one notable exception. This is particularly pleasing, since within the context of the strict dependence model [1] this is predicted. The exception is for the Maron function on the Cranfield 1400. There does not appear to be an explanation for this exceptional result.

To appreciate these experimental results more exactly we have done an analysis of the feedback sets involved and the number of queries actually entering the evaluation. Tables 7, 8 and 9 show the details. For example in Table 7 for the Cranfield 1400 collection when the cut-off is 10, only 158 queries out of 225 enter the residual ranking evaluation because for the initial co-ordination level search 49 queries do not have any relevant documents in the feedback set and 18 queries have all their relevant documents in the feedback set. What to do about the queries left out of the evaluation is a difficult question to which we do not have an easy answer. The distribution of relevant documents in the feedback sets is also interesting, it shows that feedback is mostly based on only a few relevant documents.

The alternative evaluation in Tables 10, 11 and 12, comparing a typical feedback experiment EMIM(10) with COORD(10), is to emphasise the shortcoming of the data, or perhaps the low level of effectiveness of any strategy. This is particularly important in the case of UKCIS. The tables show the number of relevant documents retrieved at different cut-off levels in the residual ranking (not to be confused with the cut-off for the feedback set), and the number of queries not retrieving any relevant documents at the same levels.

For example in the case of UKCIS I (Table 11), co-ordination level matching on the remaining documents, after removing the feedback set of 10 documents, the 62 queries entering the evaluation (see Table 8) in total retrieve only 120 relevant documents when the residual ranking is cut-off at 20. Also, at the same cut-off, 24 of these 62 queries retrieved no relevant documents at all. Compare that with the feedback strategy EMIM(10) shown in the adjoining column and one gets some idea of the dramatic improvement achieved by feedback: the number of relevant documents retrieved is doubled whereas the number of queries not retrieving any relevant documents is almost halved. However, most discouragingly, the figure at the bottom of the column shows that at a cut-off greater than 200 we still have 1807 (COORD(10)) and 1710 (EMIM(10)) relevant documents to retrieve. In other words a large number of relevant documents simply remain irretrievable whether one uses feedback or not. Probably the only way to capture these documents is through document clustering.

Concluding Remarks

We have shown how an MST derived from the distribution of co-occurrences of index terms in a document collection may be used to expand a query. The MST may be constructed using any of a number of reasonable measures of association. Within the simple feedback strategy described the different MST's lead to approximately the same retrieval effectiveness, although on the whole, the MST based on the expected mutual information measure performs marginally better than any of the others.

The method of query expansion via the MST is admittedly only very crude, but it constitutes a first step in the direction of a more refined approach. Obviously a more selective mechanism for expanding queries is needed, but this can only be done by developing some appropriate theory for following different branches of the MST. In fact one could go further than this and attempt to construct a theory which would enable a decision to be made as to whether it is more profitable to look at a nearest neighbour of a relevant document from the feedback set, or whether it is more profitable to proceed to a new search term given by the MST. No doubt ultimately some structure will be discovered which will enable, at any stage of the search, the trade-off between retrieving a nearest neighbour and selecting a closely associated search term to be evaluated.

We think that one of the major stumbling blocks to further developments in 'probabilistic information retrieval' is the lack of a comprehensive theory about the estimation, from small biased samples, of the probabilities involved. The statistical literature seems to offer little guidance on this point. We have made some ad hoc suggestions, which may well be justifiable in theoretical terms. Our experiments show that the G_{iq} weight is superior to the so-called independent weight c_i. Unfortunately this result is not unequivocal. We have found that in some rare circumstances c_i gives better performance than G_{iq} and we do not understand the reason for this. We believe that some theoretical work on the estimation rules involved may throw some light on this.

REFERENCES

1. C.J. VAN RIJSBERGEN, A theoretical basis for the use of co-occurrence data in information retrieval. J. Docum. 1977, 33, 106-119.

2. D.J. HARPER and C.J. VAN RIJSBERGEN, An evaluation of feedback in document retrieval using co-occurrence data. J. Docum. 1978, 34, 189-216.

3. S.E. ROBERTSON and K. SPARCK JONES, Relevance weighting of search terms. J. ASIS 1976, 27, 129-416.

4. C.J. VAN RIJSBERGEN, Information Retrieval, Second Edition, Butterworths, London, 1979.

5. S.E. ROBERTSON, The probability ranking principle in IR. J. Docum. 1977, 33, 294-304.

6. S.P. HARTER, A probabilistic approach to automatic keyword indexing, Ph.D. Thesis, University of Chicago, 1974.

7. V.K.M. WHITNEY, Minimal spanning tree, Algorithm 422. Commun. ACM. 1972, 15, 273-274.

8. R.C. PRIM, Shortest connection networks and some generalizations. Bell Syst. Tech. J. 1957, 36, 1389-1401.

APPENDIX

The MST's used in these experiments were generated by a BCPL program which follows the algorithm given by Whitney [7], which in turn is an encoding in FORTRAN of the 'classic' algorithm of Prim [8]. To connect N terms in an MST, the Whitney/Prim algorithm requires $O(N^2)$ term comparisons. Each term comparison involves the computation of the number of documents in which the terms co-occur, and (in the case of EMIM) the computation of four logarithms. Since N will usually be several thousand (typically between 5000 and 10000) a crude encoding of the algorithm would lead to a program that was too slow to be of any utility. But by making use of a number of optimisation strategies, an MST program was devised which, although heavy on computing resources, is practicable.

The most important optimisation strategy is worth describing. If t is a term in a document collection, we denote by D(t) the set of documents in which t occurs. If d is a document, we denote by T(d) the set of terms contained in d. The mapping d -> T(d) is given by the document file, and t -> D(t) by the inverted file. The set of those terms with which t co-occurs in at least one document, C(t), is given by

C(t) = union (T(d) ¦ d in D(t))

(i.e. the union of the sets T(d) for which d is in D(t)), which may be written

C(t) = T(D(t))

If we make the assumption that links in the MST will not be between terms with zero co-occurrence, then we do not need to compare t with any terms outside the set C(t). The program therefore estimates the size of C(t) for each t, and if it is small compared with the total term size, computes C(t) and uses this as a list of terms for comparison with t. In the case of Cranfield 1400, for example, only 10 of a potential total number of 3597903 comparisons are made to construct an MST.

Two further points need to be made. The first is that C(t) is computed by mapping t -> D(t) -> T(D(t)), and this involves easy access to the document file and inverted document file. In fact these are both held in core in our implementation, and this imposes a strict upper limit to the size of the collections for which the MST can be generated. The second point is that for certain similarity measures (EMIM is one of them) terms with co-occurrence zero can in principle be linked in the MST. Consequently the MST's used in these experiments are not necessarily exact, although we believe that the disparity, if there is one, is very slight, and should not affect the experimental results.

Cosine	$\frac{P(x_i=1,\ x_j=1)}{\sqrt{(P(x_i=1)\ P(x_j=1))}}$
Dice	$\frac{2\ P(x_i=1,\ x_j=1)}{P(x_i=1)+P(x_j=1)}$
EMIM	$\sum_{x_i,x_j} P(x_i,x_j) \log \frac{P(x_i,x_j)}{P(x_i)P(x_j)}$
Maron	$P(x_i=1,\ x_j=1) - P(x_i=1)\ P(x_j=1)$
Rajski	EMIM/Entropy
Entropy	$-\sum_{x_i,x_j} P(x_i,x_j) \log P(x_i,x_j)$

Table 1

The different association measures used to generate the MST's.

	Cranfield 1400	UKCIS I	UKCIS II
no. of documents	1400	11613	15748
no. of terms	2683	12000	8882
no. of requests	225	142	152
average no. of terms per document	29.9	6.8	6.4
average number of relevant documents per request	7.2	28.6	43.8

Table 2

Collection details.

Cutoff=10

	Cranfield 1400		UKCIS I		UKCIS II	
\ P R \	EMIM	IND	EMIM	IND	EMIM	IND
0	45.77	37.53	50.52	41.61	57.21	41.72
10	43.42	35.59	35.68	23.89	38.43	29.65
20	38.26	31.91	27.43	18.62	26.79	20.02
30	33.85	28.42	22.25	15.73	21.04	15.44
40	29.76	25.91	18.14	12.58	14.68	12.46
50	27.11	23.93	15.82	11.04	12.18	9.10
60	21.53	17.90	13.35	9.01	8.53	6.66
70	19.25	15.07	9.43	6.16	5.65	5.25
80	16.90	12.91	8.94	5.71	3.81	3.34
90	13.47	10.90	5.17	2.81	1.47	1.69
100	13.25	10.64	3.77	1.55	1.37	1.51

Cutoff=20

	Cranfield 1400		UKCIS I		UKCIS II	
\ P R \	EMIM	IND	EMIM	IND	EMIM	IND
0	41.59	33.27	43.82	36.38	57.38	36.79
10	38.48	30.60	31.55	21.03	38.32	26.92
20	33.37	27.75	24.82	16.06	25.91	17.53
30	28.14	22.58	20.94	13.60	19.95	13.79
40	24.97	20.93	16.73	11.32	15.13	10.81
50	23.28	19.28	14.76	10.11	10.29	6.80
60	16.77	13.25	9.49	6.12	7.76	5.06
70	13.85	10.12	7.39	4.66	5.60	4.12
80	12.33	8.88	6.75	4.27	4.22	2.94
90	10.66	7.66	5.19	2.52	1.46	1.55
100	10.42	7.38	4.05	1.36	1.31	1.40

Table 3

A comparison, in terms of precision and recall, of two relevance feedback strategies with query expansion via the MST based on the expected mutual information measure. EMIM uses the G_{iq} weight and IND uses the independent c_i weight. Cutoff indicates the size of the feedback set.

Cranfield 1400 (cutoff=10)

\ P	Cosine	Dice	EMIM	Maron	Rajski	Co-ord
R \						
0	42.43	43.96	45.77	47.82	44.61	28.30
10	40.21	42.25	43.42	45.58	41.59	26.73
20	35.35	37.86	38.26	39.09	38.26	24.86
30	30.45	33.82	33.85	34.83	32.83	20.62
40	27.70	29.53	29.76	31.39	28.67	17.15
50	24.95	27.72	27.11	29.27	26.53	15.12
60	19.31	21.28	21.53	22.59	20.73	10.92
70	15.13	16.25	19.25	19.43	17.24	8.97
80	13.11	14.37	16.90	16.57	15.34	7.36
90	11.10	11.87	13.47	13.29	12.47	6.13
100	10.95	11.64	13.25	13.03	12.29	6.00

Cranfield 1400 (cutoff=20)

\ P	Cosine	Dice	EMIM	Maron	Rajski	Co-ord
R \						
0	38.94	38.94	41.59	42.99	38.53	17.90
10	37.13	36.64	38.48	39.99	36.25	16.29
20	31.79	31.55	33.37	34.65	32.64	15.58
30	26.97	26.02	28.14	29.29	27.24	12.67
40	24.11	23.31	24.97	25.68	22.95	10.58
50	21.84	22.03	23.28	23.72	21.08	10.09
60	15.20	16.35	16.77	17.25	16.38	7.70
70	12.23	11.99	13.85	13.97	13.46	5.34
80	11.41	10.84	12.33	12.14	11.92	4.94
90	9.85	9.32	10.66	10.51	10.62	3.97
100	9.61	9.08	10.42	10.21	10.37	3.89

Table 4

A comparison of the effectiveness, in terms of precision and recall, of different MST's based on a range of association measures. Each experiment uses relevance feedback and expansion. Cutoff indicates the size of the feedback set.

UKCIS I (cutoff=10)

\ P	Cosine	Dice	EMIM	Maron	Rajski	Co-ord
R \						
0	49.73	47.75	50.52	48.80	48.63	30.77
10	33.09	32.61	35.68	34.03	31.98	16.73
20	25.61	24.90	27.43	25.67	24.34	11.86
30	22.67	21.94	22.25	21.25	22.01	9.79
40	18.55	17.85	18.14	17.69	18.19	7.20
50	15.93	15.18	15.82	15.30	15.82	5.77
60	13.40	12.79	13.35	13.22	13.37	4.92
70	8.71	8.60	9.43	9.27	8.68	3.29
80	8.06	7.10	8.94	8.84	8.06	1.87
90	5.87	4.97	5.17	5.25	5.87	1.18
100	4.47	3.57	3.77	3.78	4.48	0.59

UKCIS I (cutoff=20)

\ P	Cosine	Dice	EMIM	Maron	Rajski	Co-ord
R \						
0	38.82	39.94	43.82	43.37	38.24	21.04
10	29.13	27.86	31.55	30.02	28.68	11.52
20	23.75	22.15	24.82	23.94	22.72	8.22
30	20.67	18.91	20.94	19.85	19.99	6.48
40	16.45	15.76	16.73	16.28	15.69	5.28
50	14.18	13.25	14.76	14.09	13.94	4.60
60	9.36	8.39	9.49	9.58	9.37	2.92
70	7.34	6.33	7.39	7.39	7.36	2.45
80	6.85	5.93	6.75	6.79	6.87	2.24
90	5.19	4.27	5.19	5.28	5.19	1.57
100	4.05	3.12	4.05	4.05	4.05	1.16

Table 5

As for Table 4 but with a different test collection.

UKCIS II (cutoff=10)

\ P	Cosine	Dice	EMIM	Maron	Rajski	Co-ord
R \						
0	50.60	51.66	57.21	55.62	52.56	28.29
10	36.20	36.04	38.43	35.85	35.69	17.53
20	27.11	27.89	26.79	25.07	27.43	13.30
30	21.34	20.74	21.04	19.73	20.40	11.14
40	14.77	14.51	14.68	14.62	14.96	7.29
50	12.05	12.13	12.18	11.79	12.20	5.62
60	8.14	8.17	8.53	8.31	8.48	4.33
70	5.45	5.46	5.65	5.14	5.43	3.74
80	3.66	3.71	3.81	3.37	3.67	1.81
90	1.24	1.25	1.47	1.15	1.24	0.80
100	1.12	1.12	1.37	1.08	1.12	0.66

UKCIS II (cutoff=20)

\ P	Cosine	Dice	EMIM	Maron	Rajski	Co-ord
R \						
0	51.45	53.59	57.38	56.12	52.65	21.11
10	35.63	37.08	38.32	36.81	36.83	13.26
20	25.02	25.70	25.91	24.31	25.79	8.06
30	18.27	18.30	19.95	18.93	17.95	6.39
40	13.65	13.88	15.13	14.59	13.84	5.37
50	9.65	10.16	10.29	9.97	9.89	3.67
60	7.13	7.38	7.76	7.15	7.11	2.76
70	5.18	5.17	5.60	4.83	5.16	2.35
80	4.06	4.14	4.22	3.80	3.99	1.38
90	1.25	1.27	1.46	1.16	1.25	0.80
100	1.10	1.12	1.31	1.05	1.12	0.69

Table 6

As for Table 4 but with a different test collection.

Feedback set for Cranfield 1400 (cutoff = 10)

no. of queries = 225
no. of queries in evaluation = 158
no. of queries with no relevant documents = 49
no. of queries with all relevant documents = 18

Distribution

no. of rels:	1	2	3	4	5	6	7	8	9	
no. of queries:	51	35	37	17	14	1	1	1	1	(total 158)

Feedback set for Cranfield 1400 (cutoff = 20)

no. of queries = 225
no. of queries in evaluation = 164
no. of queries with no relevant documents = 32
no. of queries with all relevant documents = 29

Distribution

no. of rels:	1	2	3	4	5	6	7	8	9	10	11	
no. of queries:	39	34	31	21	15	11	3	5	1	1	3	(total 164)

Table 7

A breakdown of the feedback sets. The distribution shows the number of queries that have a different number of relevant documents in the feedback set.

Feedback set for UKCIS I (cutoff = 10)

no. of queries = 142
no. of queries in evaluation = 62
no. of queries with no relevant documents = 77
no. of queries with all relevant documents = 3

Distribution

no. of rels:	1	2	3	4	5	6	8	9	
no. of queries:	35	10	1	6	3	3	2	2	(total 62)

Feedback set for UKCIS I (cutoff = 20)

no. of queries = 142
no. of queries in evaluation = 72
no. of queries with no relevant documents = 66
no. of queries with all relevant documents = 4

Distribution

no. of rels:	1	2	3	4	5	6	7	8	10	12	18	
no. of queries:	30	12	9	4	4	2	3	3	2	2	1	(total 72)

Table 8

As for Table 7 but with a different test collection.

Feedback set for UKCIS II (cutoff = 10)

no. of queries = 152
no. of queries in evaluation = 80
no. of queries with no relevant documents = 70
no. of queries with all relevant documents = 2

Distribution

no. of rels:	1	2	3	4	5	6	8	9	10	
no. of queries:	33	16	11	7	3	2	5	1	2	(total 80)

Feedback set for UKCIS II (cutoff = 20)

no. of queries = 152
no. of queries in evaluation = 88
no. of queries with no relevant documents = 62
no. of queries with all relevant documents = 2

Distribution

no. of rels:	1	2	3	4	5	6	8	9	10	12	14	15	16	18	20	
no. of queries:	28	16	15	8	6	2	3	1	2	1	2	1	1	1	1	(total 88)

Table 9

As for Table 7 but with a different test collection.

Cranfield 1400 (cutoff = 10)

	Co-ord		EMIM	
Cutoff	Rels Retr.	Retr. None	Rels Retr.	Retr. None
10	148	74	250	41
20	234	44	337	26
30	303	30	406	18
40	351	25	460	14
50	393	24	496	12
60	428	20	533	12
70	462	18	568	11
80	489	15	598	10
90	512	14	616	10
100	526	14	635	8
110	543	13	658	7
120	561	12	671	7
130	578	11	688	6
140	593	10	703	6
150	603	10	714	6
160	611	10	720	6
170	627	10	735	6
180	652	7	746	6
190	663	6	753	6
200	679	5	758	6
200+	249		170	

Table 10

The residual ranking is cut-off at different values 10(10)200, and at each one 'Rels Retr.' indicates the number of relevant documents retrieved at that rank position. 'Retr. None' indicates the number of queries that have retrieved no relevant documents at that rank position. The rank position 200+ shows the number of relevant documents that remain to be retrieved at rank 200. Cutoff indicates the size of the feedback set.

UKCIS I (cutoff = 10)

	Co-ord		EMIM	
Cutoff	Rels Retr.	Retr. None	Rels Retr.	Retr. None
10	74	33	156	18
20	120	24	240	15
30	170	21	313	13
40	225	19	366	13
50	265	15	391	13
60	286	12	413	13
70	297	12	439	12
80	306	12	454	12
90	321	11	466	12
100	333	10	484	12
110	347	10	494	12
120	368	10	500	12
130	387	8	508	11
140	396	8	516	11
150	402	8	520	11
160	416	8	529	10
170	437	8	541	10
180	448	8	544	10
190	466	8	558	9
200	475	8	572	9
	---		---	
200+	1807		1710	

Table 11

As for Table 10 but with a different test collection.

UKCIS II (cutoff = 10)

	Co-ord		EMIM	
Cutoff	Rels Retr.	Retr. None	Rels Retr.	Retr. None
10	101	46	240	20
20	173	37	386	15
30	233	29	491	13
40	282	24	577	12
50	332	23	651	11
60	369	22	707	9
70	406	21	761	9
80	446	18	802	8
90	479	17	849	8
100	506	13	906	8
110	531	12	961	8
120	548	12	1005	8
130	567	12	1059	8
140	586	11	1098	8
150	622	9	1135	8
160	639	9	1171	8
170	671	8	1214	7
180	695	8	1250	6
190	717	7	1276	6
200	732	7	1295	6
	---		---	
200+	4538		3975	

Table 12

As for Table 10 but with a different test collection.