Information Technology: Research and Development (1983), *2* (1-21)

EXPERIMENTS WITH REPRESENTATION IN A DOCUMENT RETRIEVAL SYSTEM*

W. B. CROFT

Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003, USA

(Received 9 June 1982, revised 6 September 1982)

ABSTRACT

A series of experiments are described which demonstrate the effectiveness of incorporating term significance weights in document retrieval systems based on probabilistic models. Term significance weights are derived during the indexing process and represent a term's importance in individual documents. The weights in this paper come from within-document frequencies and they are used in search strategies based on a recently proposed modification to the probabilistic models. These search strategies provide a means of improving search performance with very little system overhead.

1. INTRODUCTION

The two main processes in a document retrieval system are *indexing* and *retrieval.* The indexing process analyzes the text of the documents to produce *representations* of their contents. These document representatives are used by the retrieval process in conjunction with a user's query to determine which documents are relevant to that query. In many previous experiments, the search strategies used in the retrieval process have been based on a probabilistic model which assumes that documents are represented by *binary* index terms (Robertson and Sparck Jones, 1976; Harper and van Rijsbergen, 1978; Sparck Jones, 1979a, b; Robertson *et al.*, 1980). Binary terms are either assigned (value 1) or not assigned (value 0) to document representatives. This paper describes a series of experiments using search strategies based on a recently proposed model which assumes that the document representatives contain weighted terms (Croft, 1981). The weights (known as *term significance weights*) measure how important a term is in describing the content of a particular document. Examples of this type of weight are the within-document frequency (Salton, 1968) and the weight proposed by Harter (Harter, 1975).

Term significance weights are derived during the indexing process and give more information about a document's content than simple binary terms. This is in contrast to weights such as the inverse document frequency (Sparck Jones and Bates, 1977) or the weights from the other probabilistic models. These weights,

* This research was supported by NSF grant IST-8011605.

0144-817X/83/01 0001-21 $03.00

which can be calculated at search time, are derived using statistics from collections of documents which have been indexed with binary terms. Term significance weights, therefore, can be characterized as using local information about word occurrences in documents whereas the other weights use more global information. The model on which this research is based was designed to make full use of both types of information in order to improve retrieval effectiveness.

Evidence of the usefulness of term significance weights when used in connection with the probabilistic models is extremely limited and not conclusive (Robertson *et al.*, 1980). The experiments in this paper were designed, therefore, to test the effectiveness of the proposed model. The next section outlines, in brief, the model with term significance weights. Sections 3 and 4 describe the test collections and the evaluation methods used. The last two sections describe the experiments in detail and summarize the results.

2. THE PROBABILISTIC MODEL

The basic probabilistic model assumes that binary terms are assigned independently to document representatives (van Rijsbergen, 1979). It shall be referred to here as the *binary independence model.* Documents are ranked according to their probability of relevance to the query. This effectively means that the documents are ranked on the basis of the following score

$$\Sigma x_i \log[p_i(1-q_i)/(1-p_i)q_i] \tag{1}$$

where x_i is the ith term in a document, p_i is the probability that term i is 1 (assigned) in the set of relevant documents, q_i is the probability that term i is 1 in the set of nonrelevant documents, and the summation is over all terms (in practice this is usually restricted to terms in the query). The probabilities p_i and q_i can be estimated using information from the relevance feedback process in which a user identifies some of the relevant documents from the top of the initial ranking. Typically this identification is done for the top 10 or 20 retrieved documents. For the initial search (prior to feedback), the same model can be used but the values of p_i will be very approximate (Croft and Harper, 1979).

The model described by Harter (1975) pointed out that the main problem in indexing is the assignment of index terms to the document representative. The frequency of an index term or word's occurrence in the document text can be used to help in this assignment process. In Harter's model, a measure of term significance was proposed and terms which had a greater level of significance than some threshold value were assigned to the document. The model used here interprets significance weights as being estimates of the probability that a term is assigned to a document. A high term significance weight implies a high probability of being assigned, a low weight implies a lower probability and a zero weight (no occurrences in the document text) implies a zero (or near zero) probability of assignment. This means that the two states of a binary term in a document ($x_i = 1$ or 0) have probabilities associated with them. Therefore, instead of ranking documents according to the score in equation (1), documents should be ranked by the *expected value* of this score. Given the properties of expected values, this new score can be shown to be

$$\Sigma P(x_i = 1 \mid d) \log[p_i(1-q_i)/(1-p_i)q_i] \tag{2}$$

where $P(x_i \mid d)$ is the probability that term i is assigned to document d. The derivation of equation (2) is discussed in detail in Croft (1981). $P(x_i = 1 \mid d)$ is estimated using term significance weights and is zero for a term which does not occur in a document. In this paper, we shall concentrate on the use of the within-document frequency w_{di} (the number of times term i occurs in document d) as the term significance weight. This is the simplest information about a term's importance in a document that can be derived from the indexing process. More sophisticated weights, such as those in the Harter model (Harter, 1975), involve many parameters that can be difficult to estimate.

Retrieval models which assume dependence between binary terms have also been proposed (van Rijsbergen, 1977). Although these models can be extended to include term significance weights (Croft, 1981), they have not been shown to have consistently better results than the model assuming independence of terms. Therefore, the experiments in this paper will use the modified independence model only.

3. TEST COLLECTIONS

The two collections of documents used are the Cranfield collection of 1400 aeronautics documents and the National Physical Laboratory collection of 11 429

Table 1. Collection statistics

Collection	*CA*	*NPL*
Number of documents	1400	11429
Number of terms	4949	7491
Average terms per document	53.6	20.0
Average documents per term	15.2	30.4
Number of queries	225	93
Average terms per query	8.9	7.2
Average relevant documents per query	7.2	22.4

Table 2. Comparison of within-document frequencies

	Percentage of occurrences	
Frequency	*CA*	*NPL*
1	70.9	84.0
2	17.3	12.5
3	6.1	2.6
4	2.6	0.7
5	1.4	0.2
6	0.7	0.06
7	0.4	s
8	0.3	s
9	0.1	s
10	0.1	s
11	0.06	—
12	s	—
13	s	s
14	s	—
15–27	s	—

* s means less than 0.01 per cent

documents in physics and computer science. The Cranfield collection (CA) was indexed automatically using the SMART routines (Salton, 1968) and the National Physical Laboratory collection (NPL) was indexed automatically using the Cambridge routines (Sparck Jones and Bates, 1977). In each case, every term in a document has a weight associated with it which is the frequency of occurrence of the term in the document text. The terms in the sample queries are treated as strictly binary in these experiments. This is done to isolate the effects of using weights derived from the document indexing.

Table 1 gives the statistics for the indexed documents, queries and relevance judgements in these collections. Table 2 gives the distribution of within-document frequencies. This table shows that the CA collection has a wider range of frequencies and a much higher percentage of frequencies greater than one. Much of this difference is due to the size of the text that was used for the indexing. For the CA collection, this was an average of over 150 words per document whereas for the NPL collection it was less than 50 words per document.

4. EVALUATION

The average performance of a given retrieval strategy is measured by using an evaluation method in conjunction with the relevance judgements. In order to avoid some of the controversy over which method should be used, we have chosen three different types of evaluation for these experiments. The first method is a precision–recall table which gives average precision values at standard recall levels. The exact method of calculation of these values is described by Harper and van Rijsbergen (1978). In some cases where detailed comparisons are not necessary, this is the only evaluation method used.

To emphasize the performance of the searches at the top end of the document rankings, the E measure (van Rijsbergen, 1979) is used as the second method. The E measure is a weighted combination of precision and recall such that the lower the E value, the greater the effectiveness. The parameter B is used to reflect the emphasis on precision or recall. $B = 1$ corresponds to attaching equal importance to precision and recall. $B = 0.5$ and 2 corresponds to attaching half and twice as much importance to recall as to precision, respectively. The E measure evaluates a set of retrieved documents. This set is defined by establishing cutoff points in the document ranking. A typical cutoff point is to take the top 10 documents. Evaluation with cutoff points has the advantage that the ranks within the retrieved set do not affect the evaluation whereas precision–recall figures can be very sensitive to the exact rankings. For example, at a cutoff of 10 documents, the E measure considers simply the number of relevant documents in this set rather than their positions in the ranked list. Since the user must examine the ten documents anyway in the typical relevance feedback process, the E measure seems appropriate.

Another advantage of the E measure is that significance tests are easy to apply because there is a single E value (for a given value of B) for each query rather than a set of recall–precision figures. The significance test which is used in these experiments is the sign test with a significance level of 0.05. Some pairs are eliminated from this test if their difference is less than a specified percentage of their value. The sign test is appropriate because it assumes only that the E measure is an ordinal scale (Siegel, 1956). Other tests, such as the Wilcoxon signed-ranks test, are more powerful (in the sense that it is easier to establish significance) but they may not be applicable. The results of the significance tests for selected pairs of strategies are

given in the tables in the appendix. The E evaluation in the experiments is presented as average E values for $B = 0.5, 1, 2$ at two cutoffs of 10 and 20 documents from the top of the document ranking.

The third evaluation method is simply to list the number of queries that do not retrieve relevant documents and the total number of relevant documents retrieved over all queries at particular cutoffs (10 and 20 documents) in the ranking.

For the experiments which use relevance feedback, a recall–precision evaluation method known as residual ranking is also used (Harper and van Rijsbergen, 1978). In this method, documents used in the relevance feedback process are removed from the final evaluation. This is done so that the document ranking after feedback can be directly compared with the documents the users would have seen if they had continued to examine the initial ranking past the top 10 or 20 documents that were used for feedback. The residual ranked feedback results are compared with the simple coordination match (documents ranked according to the number of matching terms with the query) and the upperbound feedback (where p_i, q_i are estimated using the full set of relevant documents for each query) by removing the same 10 or 20 documents used for feedback from each of the rankings.

The accuracy of the experimental retrieval system was confirmed by comparing results from this system with results obtained in previous work. In particular, the results of Croft and Harper (1979), Harper and van Rijsbergen (1978) and the coordination match experiment in Robertson *et al.* (1980) were used for this comparison.

5. EXPERIMENTS

The following subsections present the results of retrieval experiments designed to test the effectiveness of the model mentioned in section 2. Each subsection deals with a different retrieval situation, such as prior to feedback, after feedback and upperbound experiments. The later subsections examine the effectiveness of the proposed model in these situations. The results of each experiment for both test collections are shown together in Appendix A.

5.1 *Basic search strategies*

The first experiments employ three basic search strategies which are used for the initial search before feedback:

1. Ranking the documents using the coordination match.
2. The cosine correlation (Salton, 1968).
3. The inverse document frequency weight.

These results will serve as benchmarks for later experiments as well as providing a rough measure of how well the two collections respond to different types of search strategies. The coordination match is the simplest strategy as it ranks documents solely by the number of terms they have in common with the query. The inverse document frequency method assigns a weight to each matching term which measures the term's usefulness for retrieval. A document is ranked according to a score formed by summing the weights of the terms occurring in the document. The cosine correlation produces a ranking of documents which takes within-document frequencies into account. It measures the similarity between a query and a document

when both are treated as vectors in 'term space'. This weight is, therefore, one way of incorporating term significance weights in the search strategy.

Table A.1 gives the results for these experiments. Two sets of figures are given for the cosine correlation, one where the index terms are assumed to be binary (COSB) and the other where the within-document frequencies of the terms are used (COSW).

Both collections have significantly better performance when the inverse document frequency weight is used rather than the coordination match. There is a large difference, however, between the two collections when the cosine correlation is considered. In both cases the cosine correlation including within-document frequencies (COSW) outperforms the binary cosine correlation (COSB), but for the CA collection the relative performance of COSW is much better. Indeed, in this case, COSW is not significantly different from the inverse document frequency weight, whereas for the NPL collection, COSW does not perform as well as the coordination match. This shows that the heuristic use of term significance weights need not give any performance benefit. In fact, for the NPL collection, the performance definitely decreases.

5.2 *Combination match*

It has been pointed out (Croft and Harper, 1979) that the binary independence model can be used for the initial search before feedback if approximations are used for p_i. If p_i is assumed to be a constant for all terms in the query, the ranking function for the documents (derived from equation (1)) is

$$C\Sigma x_i + \Sigma x_i \log(1-q_i)/q_i \tag{3}$$

The second term of equation (3) is essentially the inverse document frequency weight and the first is simply a constant (C) times the number of matching terms in the query and document. This ranking function is known as the *combination match.* The combination match has been shown to be effective with other collections (Croft and Harper, 1979; Harper, 1980) and in Table A.2 the results for the CA and NPL collections are given. This table contains the results for the best values of $C(p_i=0.6$ for CA, 0.7 for NPL), but there is little difference for both collections in the range $p_i=0.6$ to 0.9. In neither case is there any significant difference between the combination match and the inverse document frequency weight, although, for the NPL collection, the combination match does have a slight edge in recall–precision figures and in the number of queries retrieving relevant documents. When considered with the previously mentioned experiments, this means that for all practical purposes, the inverse document frequency weight is a good choice for the initial search.

5.3 *Searches with feedback*

In this section, we shall present the results of searches which use relevance feedback information to estimate the values of p_i and q_i. The first experiment of this type is the upperbound strategy. This experiment gives the effective performance limit of strategies based on the binary independence model because it uses the complete relevance judgements for each query to estimate the parameters. The upperbound results are given in Table A.3. For both collections, very significant performance improvements over the coordination match appear to be possible.

In Table A.4, the values of actual feedback experiments are compared with the

upperbound and coordination matches. This entire table uses residual ranking for evaluation. The actual feedback experiments are done using the relevant documents in the top 10 or 20 documents from the coordination match (FEED10, FEED20) to estimate the values of p_i and q_i. This process simulates what would happen in a real system where complete relevance judgements are not obtainable. This table shows that, for both collections, the use of the binary independence model with feedback can significantly improve performance. This was also shown by Sparck Jones (1979a, b) and Robertson *et al.* (1980). In the next section we shall demonstrate how the use of term significance weights with the binary independence model can further improve performance.

One interesting outcome of this set of experiments dealt with the estimation problem and the EMIM weight (Harper and van Rijsbergen, 1978). The EMIM weight is heuristic in nature and is derived from the same information as the weight used in the binary independence model ($\log p_i(1-q_i)/(1-p_i)q_i$). In previous experiments it outperformed the theoretically superior independence weight. If the number of relevant documents in the retrieved set (10 or 20 top ranked) which contain a term x_i is r and the total number of relevant documents in the retrieved set is R, then p_i is normally estimated as $r+0.5/R+1.0$ (Robertson and Sparck Jones, 1976). However, when $r=0$, this formula gives a very bad estimate. Therefore a simple rule of using $p_i=0.01$ in this case is used. With this modification to the estimation, the normal binary independence weight performed at least as well as the heuristic EMIM weight. Therefore, the EMIM weight performs well because it avoids some estimation problems rather than by exploiting some failure of the independence model. The only disadvantage of using the new estimation is that, for the NPL collection, some relevant documents that were far down in the initial ranking were pushed even further down after feedback (note the recall = 100 entry for the NPL FEED10 and FEED20). This happens because terms that do not occur in the retrieved relevant set are given lower weights than with the previous estimation method.

5.4 *Incorporating term significance weights before feedback*

The search strategy used in this section incorporates term significance weights in the probabilistic model according to equation (2). For the initial search prior to feedback, this means that the weight calculated for each term in the summation is $P(x_i \mid d)$ multiplied by the weight calculated for the combination match ($C+\log(1-q_i)/q_i$—*see* equation 3).

$P(x_i \mid d)$ is estimated from within-document frequencies in the following way. If w_{di} is the within-document frequency of term i in document d, then n_{di}, which is the normalized within-document frequency, is calculated as $w_{di}/\max\{w_{d1}, w_{d2}, \ldots\}$. $P(x_i \mid d)$ is then estimated as $K+(1-K)n_{di}$, where K is a constant between 0 and 1. When K is 0, $P(x_i \mid d)$ is estimated by the normalized within-document frequency. For example, a term that occurs once in a document which has a maximum within-document frequency of 5 has a value of $1/5=0.2$ for $P(x_i \mid d)$. This is a low estimate for a term which is actually in the document. It must be remembered that term occurrence in a document is a rare event in that only very few terms out of a large possible vocabulary occur in any given document. This implies that any non-zero significance weight should give a reasonably high estimate for the probability of assignment. The constant K is introduced, therefore, to give higher estimates for these probabilities. If $K=0.5$, the estimates for $P(x_i \mid d)$ will range from 0.5 to 1 or if

$K=0.9$, the estimates range from 0.9 to 1. The experiments in this section and in section 5.5 will test the effect of varying K on the retrieval performance.

Table A.5 gives the results of using the estimated values of $P(x_i \mid d)$ in the new model. The search names indicate the value of K used for the estimation, for example, NEW7 is the new model with $K=0.7$. For comparison, the search NEWF, which uses the unnormalized within-document frequency in the new model, is also presented.

It is clear from these figures that the unnormalized within-document frequency is a very poor estimator for $P(x_i \mid d)$. In fact, it decreased performance relative to the combination match. We should expect this result as a frequency looks nothing like a probability, but this is one example of how a search strategy based on a model can show how information should be used. When the normalized values of the within-document frequencies are used, the value of K does significantly affect performance. For example, with $K=0$ the new search does significantly better than the combination match for the CA collection, but significantly worse for the NPL collection. For higher values of K, the new search is significantly better than the combination match and the inverse-document frequency weight in both collections. Differences in the recall–precision figures between the new search and the other strategies ranged between 10 and 20 per cent for the CA collection and up to 10 per cent for the NPL collection. In both cases, the sign test confirmed that the differences were significant. The optimum value of K was different for the two collections (0.3 for CA, 0.5 for NPL), but choosing K to be 0.5 seems to be reasonable if no tuning is possible.

Two other experiments suggested by Croft (1981) were also carried out. The first used the within-document frequency weights in an attempt to get better estimates of p_i and q_i than the binary terms provide. The second experiment was to use the Edmundson and Wyllys weight (1961) to estimate $P(x_i \mid d)$. Both of these experiments resulted in very poor performance so these techniques were not pursued further.

5.5 *Incorporating term significance weights after feedback*

The next set of experiments tested the effect of incorporating $P(x_i \mid d)$ in the upperbound and relevance feedback searches according to equation (2). The estimation of this probability is done the same way as in the last section. Table A.6 shows the results of the new model for the upperbound experiment. The unmodified upperbound is included for comparison. The values of K used in the estimation are again included in the search title (e.g., UPPER3 is the new upperbound with $K=0.3$).

It can be seen that, for both collections, incorporating the term significance weights significantly improves the potential performance. Recall–precision figures increase by up to 10 per cent in the case of the CA collection and up to 5 per cent in the case of the NPL collection. The sign test confirmed that the differences are significant. It is interesting to note that the optimum value of K has changed from the last section (from 0.3 to 0.5 for CA; from 0.5 to 0.7 for NPL). The reason for this is not known.

Table A.7 gives the results (evaluated using residual ranking) of relevance feedback searches including term significance weights. The original feedback experiments have been included for comparison. For both collections and both cutoff points (using 10 and 20 documents for feedback), the new search strategy leads to very significant improvements. For the CA collection using 10 documents for

feedback, the recall–precision figures increase by as much as 35 per cent. The NPL collection has less dramatic improvements (up to 10 per cent), but they are still significant. The optimum value of K is essentially the same as for the upperbound experiment.

6. SUMMARY OF RESULTS

The experimental results presented in section 5 lead to a single major conclusion—the use of search strategies based on the new model incorporating term significance weights will significantly improve the performance of a document retrieval system. This statement was shown to be true for two test collections which are completely different, both in their sizes and in their statistical properties. It was also shown to be true in three different retrieval situations: before feedback, after feedback and the hypothetical upperbound case where all relevant documents are known.

The term significance weights were estimated using within-document frequencies which can be derived simply and efficiently in the indexing process. Other possibilities, such as the Harter weights, have been disappointing in performance (Robertson *et al.*, 1980) due to the number of parameters that must be estimated. The only parameter which must be specified in the estimation used here is the value of K. Although this could be tuned for a given collection, it appears that a value between 0.3 and 0.7 will always give reasonable performance.

The importance of basing the search strategies on the modified probabilistic model was emphasized by the poor performance of the cosine correlation on the NPL collection. The cosine correlation does use term significance weights to do the document ranking but they are obviously used much more effectively in the new strategies.

The major overhead involved in implementing a document retrieval system based on the model tested here would be the storage of the within-document frequencies in the inverted file of documents and terms (Croft and Ruggles, 1982). However, this has been considered before in the context of using the cosine correlation and Murray (1972) discusses efficient ways of storing this information.

APPENDIX A: THE EXPERIMENTAL RESULTS

Table A.1. The coordination match (COORD), inverse document frequency (INVWT) and cosine (COSB and COSW) searches

Collection: CA

Recall	*Precision* (225 queries)			
	COORD	*INVWT*	*COSB*	*COSW*
10	40.8	47.0	44.0	47.4
20	33.7	40.9	37.5	40.3
30	26.8	33.6	30.3	31.8
40	22.4	28.9	24.4	27.2
50	20.1	26.2	21.8	23.8
60	13.4	18.5	14.2	17.8
70	10.4	13.5	10.8	13.4
80	8.6	11.3	8.8	11.5
90	7.1	8.6	7.0	8.5
100	6.7	8.2	6.6	8.1

Table A.1—*continued*

Average E values (225 queries)

Search	*Cutoff 10*			*Cutoff 20*		
	B = 0.5	1.0	2.0	0.5	1.0	2.0
COORD	84.1	82.4	78.8	88.1	85.1	78.7
INVWT	80.6	78.6	74.3	85.6	82.0	74.5
COSB	82.9	81.2	77.7	86.8	83.4	76.3
COSW	80.7	78.9	74.9	85.4	81.7	73.9

Relevance information (225 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
COORD	60	346	43	475
INVWT	47	421	36	577
COSB	60	373	33	529
COSW	52	421	34	586

Significance tests: sign test with *E* values, $B = 1$, cutoff = 10. Pairs with less than 5 per cent difference in value are ignored.

INVWT > COORD, COSB
COSW > COSB, COORD
COSB > COORD

Collection: NPL

Recall — *Precision* (93 queries)

Recall	*COORD*	*INVWT*	*COSB*	*COSW*
10	49.1	54.1	36.9	47.8
20	37.6	42.0	27.2	34.8
30	30.6	35.1	21.6	24.5
40	25.0	29.1	18.0	19.9
50	20.4	23.1	14.6	15.7
60	13.2	18.1	11.0	11.1
70	10.6	14.5	7.6	8.3
80	7.2	10.4	5.6	6.0
90	4.7	6.3	4.0	4.0
100	2.2	3.2	1.9	2.5

Average E values (93 queries)

Search	*Cutoff 10*			*Cutoff 20*		
	B = 0.5	1.0	2.0	0.5	1.0	2.0
COORD	78.6	82.6	84.2	78.3	79.3	78.3
INVWT	74.7	79.0	80.4	76.5	77.2	75.6
COSB	83.2	86.2	87.2	83.8	84.4	83.5
COSW	79.7	83.4	84.5	81.3	81.9	80.7

Relevance information (93 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
COORD	19	264	11	439
INVWT	11	306	7	472
COSB	23	206	15	328
COSW	15	249	8	377

Table A.1—*continued*

Significance tests:
INVWT > COSW, COSB, COORD
COORD > COSW, COSB
COSW > COSB

Table A.2. The combination match (COMB)

Collection: CA

Recall	*Precision* (225 queries) *COMB*
10	47.2
20	40.5
30	33.1
40	28.4
50	25.9
60	18.4
70	13.4
80	11.2
90	8.7
100	8.3

Average E values (225 queries)

Search		*Cutoff 10*			*Cutoff 20*		
	$B=$	0.5	1.0	2.0	0.5	1.0	2.0
COMB		80.8	78.9	74.7	85.6	82.0	74.4

Relevance information (225 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
COMB	48	416	36	577

Significance tests:
COMB > COSB, COORD

Collection: NPL

Recall	*Precision* (93 queries) *COMB*
10	55.8
20	44.7
30	36.8
40	30.2
50	23.7
60	17.7
70	14.5
80	10.3
90	6.4
100	3.1

Table A.2—*continued*

Average E values (93 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
COMB	74.4	79.0	80.6	76.3	77.1	75.7

Relevance information (93 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
COMB	9	312	5	477

Significance tests:
COMB > COSW, COSB, COORD

Table A.3. The upperbound experiments (UPPERB)

Collection: CA

Recall	*Precision* (225 queries) *UPPERB*
10	70.6
20	64.2
30	54.3
40	48.9
50	44.3
60	34.6
70	26.8
80	21.5
90	16.6
100	15.6

Average E values (225 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
UPPERB	71.4	68.4	62.2	79.3	74.0	62.8

Relevance information (225 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
UPPERB	18	620	12	826

Table A.3—*continued*

Collection: NPL

Recall	*Precision* (93 queries) *UPPERB*
10	69.9
20	59.8
30	52.8
40	45.5
50	38.2
60	29.4
70	23.1
80	16.9
90	11.0
100	5.7

Average E values (93 queries)

Search		*Cutoff 10*			*Cutoff 20*		
	B =	0.5	1.0	2.0	0.5	1.0	2.0
UPPERB		66.7	72.0	73.5	68.1	69.1	67.0

Relevance information (93 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
UPPERB	3	400	1	642

Table A.4. Feedback experiments. Residual ranked feedback (FEED10, FEED20) compared with residual ranked coordination match (COORD1, COORD2) and upperbound (UPPER1, UPPER2). First results use top 10 documents for feedback

Collection: CA

Recall	*Precision* (150 queries) *COORD1*	*FEED10*	*UPPER1*
10	23.1	32.8	53.3
20	17.6	28.4	45.6
30	13.1	23.0	38.8
40	10.6	20.9	34.8
50	9.0	18.5	31.9
60	6.2	13.0	25.1
70	4.4	9.6	18.6
80	3.8	7.5	14.8
90	3.1	6.5	11.8
100	2.8	6.0	11.0

Average E values (150 queries)

Search		*Cutoff 10*			*Cutoff 20*		
	B =	0.5	1.0	2.0	0.5	1.0	2.0
COORD1		92.4	91.7	90.2	92.9	91.0	86.9
FEED10		87.7	85.8	82.2	91.1	88.4	82.5
UPPER1		78.2	75.4	69.5	83.7	79.2	69.6

Table A.4—*continued*

Relevance information (150 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
COORD1	80	111	52	191
FEED10	71	173	57	232
UPPER1	25	312	16	432

Significance tests:
FEED10 > COORD1

Collection: NPL

Recall	*Precision* (74 queries)		
	COORD1	*FEED10*	*UPPER1*
10	31.9	44.7	56.6
20	25.5	37.6	50.5
30	22.0	27.9	40.6
40	17.8	22.3	33.2
50	14.8	19.5	28.2
60	11.0	16.2	22.5
70	8.6	12.4	16.4
80	7.0	9.3	13.0
90	4.5	5.0	7.8
100	2.5	2.2	3.4

Average E values (74 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
COORD1	84.9	88.1	89.4	83.7	84.7	84.4
FEED10	79.2	83.4	85.0	78.8	80.1	79.5
UPPER1	72.3	77.4	79.3	73.0	74.1	72.6

Relevance information (74 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
COORD1	21	152	15	267
FEED10	18	211	12	348
UPPER1	8	269	4	435

Significance tests:
FEED10 > COORD1

Table A.4—*continued*

(top 20 documents used for feedback)

Collection: CA			
Recall	*Precision* (165 queries)		
	COORD2	*FEED20*	*UPPER2*
10	16.0	33.5	48.5
20	13.4	28.4	42.9
30	10.1	23.8	37.7
40	8.3	19.8	31.2
50	7.3	18.7	30.0
60	5.3	11.8	21.8
70	4.3	8.5	17.5
80	3.8	7.4	15.3
90	2.6	6.5	11.9
100	2.6	6.4	11.7
Collection: NPL			
Recall	*Precision* (80 queries)		
	COORD2	*FEED20*	*UPPER2*
10	23.6	41.5	55.4
20	18.7	30.4	44.6
30	15.5	23.9	37.5
40	12.8	19.6	31.8
50	10.9	16.9	27.4
60	8.7	13.2	20.3
70	7.0	10.3	14.9
80	5.4	7.7	10.8
90	3.7	4.3	6.7
100	2.2	2.0	3.2

Table A.5. The new model. Estimates of term significance weights incorporated into the probabilistic model before feedback. NEWF uses the unnormalized within-document frequency

Collection: CA					
Recall	*Precision* (225 queries)				
	NEWF	*NEWO*	*NEW1*	*NEW3*	*NEW5*
10	45.9	49.9	51.8	53.8	53.6
20	38.4	44.5	45.3	47.4	47.1
30	30.9	36.3	37.7	40.2	39.4
40	26.2	31.5	33.5	35.3	34.1
50	23.3	28.7	29.9	31.9	31.1
60	15.5	20.7	21.7	23.1	22.0
70	12.1	15.9	16.8	17.6	16.4
80	9.6	13.1	13.7	14.1	13.3
90	7.0	9.4	9.7	10.2	9.7
100	6.6	9.0	9.2	9.6	9.1

Table A.5—*continued*

Average E values (225 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
NEWF	82.4	80.6	76.9	86.0	82.5	75.0
NEW0	79.0	77.0	72.7	83.9	79.9	71.5
NEW1	77.7	75.6	71.1	83.7	79.6	71.2
NEW3	77.5	75.3	70.6	83.4	79.1	70.4
NEW5	78.1	75.9	71.3	83.6	79.4	70.8

Relevance information (225 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
NEWF	55	385	36	560
NEW0	48	458	26	644
NEW1	47	487	25	652
NEW3	39	491	21	667
NEW5	41	476	26	659

Significance tests:
NEW0 > INVWT
NEW3 > NEW0, INVWT

Collection: NPL

Recall	*Precision* (93 queries)				
	NEWF	*NEW0*	*NEW3*	*NEW5*	*NEW7*
10	43.0	46.2	55.1	59.1	58.7
20	31.5	38.0	46.1	48.4	47.3
30	24.5	27.8	36.4	39.7	38.8
40	21.2	22.7	29.8	33.2	31.8
50	17.2	18.3	23.9	25.8	25.0
60	13.0	13.7	17.8	19.4	18.9
70	9.9	10.1	13.4	14.4	14.5
80	8.2	6.8	9.8	10.9	11.0
90	5.8	4.5	6.4	6.7	6.6
100	3.6	2.6	3.2	3.3	3.2

Average E values (93 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
NEWF	80.7	83.7	84.3	81.6	82.1	80.6
NEW0	78.8	82.3	83.3	80.2	80.9	79.6
NEW3	74.6	78.8	80.0	76.1	77.0	75.6
NEW5	73.4	77.9	79.3	75.1	76.0	74.5
NEW7	73.5	77.9	79.3	75.1	76.0	74.5

Table A.5—*continued*

Relevance information (93 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
NEWF	16	231	8	370
NEW0	12	259	11	402
NEW3	9	306	9	485
NEW5	9	322	9	503
NEW7	10	320	8	503

Significance tests:
INVWT > NEW0
NEW5 > INVWT

Table A.6. The new model. Estimates of term significance weights incorporated into the upperbound experiment

Collection: CA

Recall *Precision* (225 queries)

Recall	*UPPERB*	*UPPER0*	*UPPER3*	*UPPER5*	*UPPER7*
10	70.6	60.8	69.8	72.0	72.8
20	64.2	54.3	63.7	66.4	66.7
30	54.3	44.5	54.3	58.1	58.6
40	48.9	39.4	49.5	53.9	53.5
50	44.3	35.7	44.4	48.7	48.6
60	34.6	27.1	34.3	37.0	37.5
70	26.8	21.8	27.9	29.2	28.4
80	21.5	17.8	22.3	23.2	22.7
90	16.6	12.8	16.6	17.6	17.3
100	15.6	12.2	15.8	16.7	16.3

Average E values (225 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
UPPERB	71.4	68.4	62.2	79.3	74.0	62.8
UPPER0	74.7	72.0	66.4	81.0	76.0	65.6
UPPER3	70.3	67.3	60.8	78.5	72.9	61.4
UPPER5	69.6	66.4	59.8	78.0	72.4	60.7
UPPER7	69.9	66.8	60.3	78.4	72.8	61.2

Relevance information (225 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
UPPERB	18	620	12	826
UPPER0	27	550	14	761
UPPER3	13	646	4	861
UPPER5	9	661	4	879
UPPER7	11	655	6	864

Significance tests:
UPPER5 > UPPERB

Table A.6—*continued*

Collection: NPL

Recall	*Precision* (93 queries) *UPPERB*	*UPPER0*	*UPPER5*	*UPPER7*	*UPPER9*
10	69.9	59.0	72.0	74.5	71.8
20	59.8	46.0	60.9	62.7	61.8
30	52.8	34.2	52.8	54.9	54.2
40	45.5	28.5	44.6	46.3	45.9
50	38.2	23.6	38.0	40.4	39.6
60	29.4	18.1	29.8	30.5	30.0
70	23.1	12.8	23.0	23.7	24.0
80	16.9	9.6	17.7	18.2	18.0
90	11.0	6.3	11.4	11.9	11.9
100	5.7	4.0	6.7	6.5	6.3

Average E values (93 queries)

Search		*Cutoff 10*			*Cutoff 20*		
	B=	0.5	1.0	2.0	0.5	1.0	2.0
UPPERB		66.7	72.0	73.5	68.1	69.1	67.0
UPPER0		75.5	79.6	80.7	76.2	77.1	75.6
UPPER5		66.9	72.2	73.7	69.3	70.4	68.3
UPPER7		65.3	70.8	72.3	67.4	68.4	66.0
UPPER9		65.4	71.1	72.7	67.1	68.0	65.8

Relevance information (93 queries)

Search	*Cutoff 10* *No. queries that fail*	*Total relevant retrieved*	*Cutoff 20* *No. queries that fail*	*Total relevant retrieved*
UPPERB	3	400	1	642
UPPER0	12	301	6	485
UPPER5	6	397	3	622
UPPER7	4	416	1	655
UPPER9	3	415	1	660

Significance tests:
UPPER7 > UPPERB

Table A.7. The new model. Term significance weights incorporated into the feedback experiment. Both 10 and 20 documents used for feedback. Example: FD1N5 means feedback with 10 documents and term significance weights estimated with $K = 0.5$

Collection: CA

Recall	*Precision* (150 queries)				
	FEED10	*FD1N0*	*FD1N3*	*FD1N5*	*FD1N7*
10	32.8	42.0	42.6	43.5	42.2
20	28.4	34.5	37.6	38.1	37.3
30	23.0	26.0	28.4	29.8	29.5
40	20.9	22.8	24.2	25.6	25.3
50	18.5	20.3	22.1	23.0	22.6
60	13.0	15.7	16.2	15.9	15.8
70	9.6	11.9	11.8	11.8	11.5
80	7.5	9.0	8.9	8.8	8.6
90	6.5	7.4	7.2	7.3	7.2
100	6.0	7.1	6.8	6.8	6.6

Average E values (150 queries)

Search		*Cutoff 10*			*Cutoff 20*		
	$B=$	0.5	1.0	2.0	0.5	1.0	2.0
FEED10		87.7	85.8	82.2	91.1	88.4	82.5
FD1N0		84.4	82.6	79.0	88.3	85.2	78.6
FD1N3		83.4	81.6	78.1	87.2	83.8	76.4
FD1N5		82.9	81.1	77.3	86.9	83.4	76.0
FD1N7		83.8	82.0	78.4	87.5	84.1	77.0

Relevance information (150 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
FEED10	71	173	57	232
FD1N0	43	226	32	311
FD1N3	47	240	28	341
FD1N5	46	246	31	349
FD1N7	52	233	33	334

Significance tests:
FD1N5 > FEED10

(10 documents used for feedback)

Collection: NPL

Recall	*Precision* (74 queries)				
	FEED10	*FD1N0*	*FD1N5*	*FD1N7*	*FD1N9*
10	44.7	40.1	49.2	49.2	49.0
20	37.6	27.7	40.2	41.3	40.9
30	27.9	19.8	26.6	28.6	28.8
40	22.3	15.5	21.0	22.9	23.0
50	19.5	12.5	18.5	19.8	20.0
60	16.2	9.6	15.3	16.5	16.8
70	12.4	7.0	11.8	12.4	12.9
80	9.3	4.9	9.1	9.1	9.8
90	5.0	3.1	4.8	5.1	5.1
100	2.2	1.9	2.2	2.2	2.2

Table A.7—*continued*

Average E values (74 queries)

Search	*Cutoff 10*			*Cutoff 20*		
B =	0.5	1.0	2.0	0.5	1.0	2.0
FEED10	79.2	83.4	85.0	78.8	80.1	79.5
FD1N0	83.0	86.7	88.2	83.5	84.7	84.6
FD1N5	78.1	82.7	84.6	79.0	80.6	80.4
FD1N7	77.3	81.9	83.7	77.8	79.2	78.6
FD1N9	76.4	81.1	82.9	77.2	78.7	78.2

Relevance information (74 queries)

Search	*Cutoff 10*		*Cutoff 20*	
	No. queries that fail	*Total relevant retrieved*	*No. queries that fail*	*Total relevant retrieved*
FEED10	18	211	12	348
FD1N0	18	175	14	276
FD1N5	14	221	12	348
FD1N7	13	228	10	366
FD1N9	13	236	9	375

Significance tests:
FD1N9 > FEED10

(20 documents used for feedback)

Collection: CA

Recall	*Precision* (165 queries)			
	FEED20	*FD2N3*	*FD2N5*	*FD2N7*
10	33.5	40.3	40.0	39.7
20	28.4	33.8	33.3	33.2
30	23.8	27.6	28.4	28.4
40	19.8	21.6	22.6	23.0
50	18.7	20.3	21.1	21.1
60	11.8	13.1	13.0	12.6
70	8.5	10.0	9.8	9.6
80	7.4	8.5	8.4	8.1
90	6.5	7.0	6.8	6.5
100	6.4	6.8	6.7	6.3

Significance tests:
FD2N5 > FEED20

Collection: NPL

Recall	*Precision* (80 queries)			
	FEED20	*FD2N5*	*FD2N7*	*FD2N9*
10	41.5	43.5	45.5	46.1
20	30.4	31.5	32.4	34.1
30	23.9	21.8	24.4	25.3
40	19.6	18.3	19.4	20.1
50	16.9	15.5	16.8	17.2
60	13.2	12.8	13.5	13.9
70	10.3	10.3	10.5	10.4
80	7.7	7.8	7.8	7.8
90	4.3	4.2	4.3	4.3
100	2.0	2.0	2.0	1.9

ACKNOWLEDGEMENT

The author wishes to acknowledge the efforts of Lynn Ruggles in programming the experimental system and carrying out the experiments.

REFERENCES

Croft, W. B. (1981) Document representation in probabilistic models of information retrieval. *Journal of the American Society for Information Science 32*, 451–457.

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

Croft, W. B. and Ruggles, L. (1982) The implementation of a document retrieval system. *Proceedings of the ACM/BCS SIGIR conference*, Berlin, West Germany.

Edmundson, H. P. and Wyllys, R. E. (1961) Automatic abstracting and indexing survey and recommendations. *Communications of the ACM 4*, 226–334.

Harper, D. J. (1980) *Relevance feedback in document retrieval systems: An evaluation of probabilistic strategies.* Ph.D. Thesis, Computer Laboratory, University of Cambridge, UK.

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

Harter, S. P. (1975) A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. Part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science 26*, 197–206, 280–289.

Murray, D. M. (1972) *Document retrieval based on clustered files.* Ph.D. Thesis, Cornell University (Report ISR-20 to the National Science Foundation and to the National Library of Medicine).

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

Robertson, S. E., van Rijsbergen, C. J. and Porter, M. F. (1980) Probabilistic models of indexing and searching. *Proceedings of the ACM/BCS SIGIR conference*, Cambridge, UK.

Salton, G. (1968) *Automatic Information Organization and Retrieval.* New York: McGraw-Hill.

Siegel, S. (1956) *Non Parametric Statistics for the Behavioral Sciences.* Tokyo: McGraw-Hill Kogakusha Ltd.

Sparck Jones, K. (1979a) Experiments in relevance weighting of search terms. *Information Processing and Management 15*, 133–144.

Sparck Jones, K. (1979b) Search term relevance weighting given little relevance information. *Journal of Documentation 35*, 30–48.

Sparck Jones, K. and Bates, R. G. (1977) *Research on automatic indexing 1974–1976.* 2 vols. British Library Research and Development Report 5464, University of Cambridge, UK.

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.* 2nd edn. London: Butterworths.