Journal of the American Society for Information Science
Volume 27 number 3, May-June 1976, pp 129-146

Relevance Weighting of Search Terms

This paper examines statistical techniques for exploiting relevance information to weight search terms. These techniques are presented as a natural extension of weighting methods using information about the distribution of index terms in documents in general. A series of relevance weighting functions is derived and is justified by theoretical considerations. In particular, it is shown that specific weighted search methods are implied by a general probabilistic theory of retrieval. Different applications of relevance weighting are illustrated by experimental results for test collections.

S.E. Robertson
School of Library, Archive and Information Studies
University College London
London WC1E 6BT, England

K. Sparck Jones
Computer Laboratory
University of Cambridge
Cambridge CB2 3QG, England

• Introduction

In this paper, we examine statistical techniques for exploiting relevance information to weight search terms. The object of the paper is to find a theoretical framework which will give us some guidance as to how to make use of relevance information in searching and to test experimentally any guidance that we may find. In that event, a general probabilistic theory of relevance weighting implies that we should use a specific weighted search method and suggests a series of relevance weighting functions. The experimental results confirm some important conclusions of the theory.

• Statistical Weighting

Search term weighting is an established practice. It may be adopted simply as a means of simulating Boolean searching [see Angione (*1*)]. It is also a retrieval device in its own right. That is, term weighting may be used in a manner not equivalent to Boolean searching because retrieved documents are further ranked. We are concerned here with this more general use of weighting.

Term weights may be assigned for different reasons. One ground for weighting is user preference. The user may be more interested in documents with term *a* than documents with term *b* for reasons not directly connected with the actual use of *a* and *b* in the set of documents being searched. Request terms may also be weighted statistically (usually rather trivially) by their within-request frequency. Such user-oriented request term weights may be contrasted with document-oriented weights, and specifically with system-oriented weights reflecting the behavior of terms in the document collection as a whole. (Term weights related to individual documents, whether derived intuitively or statistically, are not explicit search term weights.)

A natural source of system-oriented term values is distributional information indicating term frequencies; and since the main problem in retrieval is to select a few relevant documents from many non-relevant ones, the general object of statistical weighting schemes is to assign high values to discriminating terms. One such scheme has been studied by Salton (*2*), and another much simpler one by Sparck Jones (*3, 4, 5*). Comparable improvements in retrieval performance have been obtained with them. In both schemes terms with medium to low collection frequencies are assigned high weights as good discriminators, while frequent terms have low weights. For term t, given

N = the number of documents in the collection, and
n = the number of documents indexed by t,
Sparck Jones assigns a weight by the function:

$$w = -\log \frac{n}{N} \ . \qquad \text{(F0)}$$

Salton's weighting scheme is very much more complicated. In retrieval, documents are ordered by notional coordination level representing the sum of matching term weights. Sparck Jones' tests with three collections showed material improvements in performance, measured by recall and precision, over unweighted terms. Statistical information, however valuable, is not readily manipulated manually, but it is well suited to automatic systems. Sparck Jones' function in particular is very simply applied at search time. Apparent merit and easy computation taken together suggest that as much distributional information as possible should be exploited in an automatic system.

The schemes described so far use only information about the distribution of terms in documents. Robertson (*6*) has drawn attention to the natural development, which is to use any available information about the distribution of terms in relevant documents. In this case term weighting becomes strictly request (and need) specific. In the previous case, a term would have the same weight in different requests. When relevant documents are taken into account, the same term may have a different value for different requests. An obvious weighting function, derived from rather different starting points, has been proposed by Barkla (*7*) and by Miller (*8, 9*). For a given term t and a given query q, if

R = the number of relevant documents for q, and
r = the number of relevant documents indexed by t,

$$w = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n}{N}\right)} \ . \qquad \text{(F1)}$$

For Barkla, an SDI service would provide relevance information via feedback; for Miller, it would be estimated by the user. Barkla's test with this weighting scheme is very difficult to interpret (see Robertson). In Miller's experiments, however, it led to a better retrieval performance for MEDLARS than standard Boolean searching. In recent tests with several collections, Sparck Jones (*10*) has shown that the function, when applied predictively as originally intended, has some merit. She has also shown that the optimal performance, which is obtained when perfect relevance information is available, *e.g.*, for test collections, is good; and she has suggested that it can be used as a general experimental yardstick. A simpler weighting scheme with the same object, using $w = r/n$, was found to be of some use by Barker, Veal and Wyatt (*11*).

Yu and Salton (*12*) have recently investigated a function defined by

$$w = \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n-r}{N-n-R+r}\right)} \ .$$

They use this function not directly as a weighting function, but in a rather complex way to modify the output of a simple coordination level matching scheme. They then *prove* that this modification of coordination level matching can be expected to improve performance. The proof involves some "independence assumptions" about the occurrence of terms which are discussed later. Although their particular weighting scheme is idiosyncratic, and their subsequent retrieval tests to confirm the proof empirically were very limited, the idea of attempting a formal proof of the validity of a particular scheme is a good one and forms a major part of the present paper.

Taken together, these results provide some *prima facie* evidence for the potential value of relevance information in statistical weighting schemes suited to automatic post coordinate term retrieval. But there is no consensus on exactly how relevance information should be exploited. What is required is a systematic theoretical and experimental investigation of the various possible ways on using relevance information for weighting search terms. This is what this paper aims to provide.

In the next section we present a series of relevance weighting functions. The theoretical framework within which these functions are derived is summarized in an informal way. A formal account of the theory is presented in full in the Appendix, to which the technically-minded reader is referred. The summary given in the text is designed to bring out the main points of the theory, justifying the particular choice of weighting function adopted and to link the theory and the experiments which are subsequently described. Thus, the theory is essentially concerned with probabilistic methods for ranking search output, to maximize recall and minimize fallout, based on assumptions about term distributions and principles of output ordering. In the experiments the methods are interpreted as request term weighting schemes, both to link the approach adopted with past work and to permit the application of a standard retrieval test methodology.

• Relevance Weighting

We assume binary index descriptions of documents. (These are more frequently encountered than non-binary ones.) We also assume a set of relevance judgments for each request. Documents may be judged relevant either in relation to the request as stated or in relation to some underlying need. In either case we hope to optimize retrieval of these or similar documents. Thus, if the judgments are specific to an individual user, the optimal retrieval strategy is to be supplied for that user. Another user, with the same verbal request but different judgments, may have a different strategy. In the paper, "request" implies specific individual need.

Now given, as above, for term t and request q,

N = the number of documents in the collection,
R = the number of relevant documents for q,
n = the number of documents having t, and
r = the number of relevant documents having t,

consider the contingency table of document distribution for t:

		Document Relevance		
		+	-	
Document Indexing	+	r	n-r	n
	-	R-r	N-n-R+r	N-n
		R	N-R	N

Relevance weighting formulae must in some way reflect the relative distribution of terms with respect to relevant and other documents. Specifically, we can derive formulae from the previous table as follows (the use of logs in all the formulae is explained in detail later):

$$w^1 = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n}{N}\right)} \qquad \text{(F1)}$$

$$w^2 = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n-r}{N-R}\right)} \qquad \text{(F2)}$$

$$w^3 = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n}{N-n}\right)} \qquad \text{(F3)}$$

$$w^4 = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n-r}{N-n-R+r}\right)} \qquad \text{(F4)}$$

Informally, for the given request term t, Function F1 represents the ratio of the proportion of relevant documents in which t occurs to the proportion of the entire collection in which it occurs, while F2 represents the ratio of the proportion of relevant documents to that of non-relevant documents. F3 represents the ratio between the "relevance odds" for the term (*i.e.* the ratio between the number of relevant documents in which it does occur and the number in which it does not occur) and the "collection odds" for t, while F4 represents the ratio between the term's relevance odds and its "non-relevance odds." Thus, the Functions F1 and F2 are related by using proportions, while F3 and F4 use odds; but F1 and F3 respectively are related by comparing the relevant document distribution of a term to its entire collection distribution, while F2 and F4 are related by comparing relevant and non-relevant distributions.

The consequences of applying these formulae, together with simple collection frequency weighting (F0), are illustrated in Table 1 and Fig. 1. Five terms are chosen: all four combinations of low and high collection frequency and low and high relevance frequency, and a "medium" term. The relationships between the weights given by each function to the different terms are shown diagrammatically in the figure (since only the ratios of the weights matter within a given scheme, all the schemes are scaled to give the same weight to the medium term e). All four functions separate the terms in the obvious ways: $b > a$ and $d > c$ for relevance frequency, $a > c$ and $b > d$ for collection frequency; in fact, all four functions give a negative weight to c (since a document chosen at random from those containing the term c is *less* likely to be relevant than one chosen at random from the whole collection). But the exact quantitative relationships are somewhat different for each function. The relationship of a and e shows that the four functions do not necessarily rank terms in the same order.

• Foundations of Relevance Weighting

So far, the weighting functions have been characterized in a fairly superficial way. In fact, all four functions derive from a formal probabilistic theory of relevance weighting. This theory is presented in full in the Appendix to which, as indicated earlier, the reader interested in its formal development or finer points is referred. This section is intended to give an informal account and interpretation of the main ideas involved.

The object of the theory is to derive an optimal ranking of the documents in a collection, on the basis of the presence or absence in each document of the request terms, when some information about the average performance of these terms is available. The theory makes use of two kinds of assumptions: *independence assumptions* and *ordering principles*.

The independence assumptions allow us to make inferences about a document containing a given combina-

Table 1. Illustrative weights for items with different distributions.

N = 200	Term *a*	*n* = 5	*r* = 1
R = 5	*b*	*n* = 5	*r* = 4
	c	*n* = 100	*r* = 1
	d	*n* = 100	*r* = 4
	e	*n* = 20	*r* = 3

Contingency table					**Weighting**				
		Document Relevance			F1	F2	F3	F4	(F0)
Document Indexing		+	−						
Term *a*	+	1	4	5	.90	.99	.99	1.08	1.60
	−	4	191	195					
		5	195	200					
Term *b*	+	4	1	5	1.51	2.19	2.19	2.89	1.60
	−	1	194	195					
		5	195	200					
Term *c*	+	1	99	100	-.40	-.40	-.60	-.62	.30
	−	4	96	100					
		5	195	200					
Term *d*	+	4	96	100	.20	.21	.60	.62	.30
	−	1	99	100					
		5	195	200					
Term *e*	+	3	17	20	.78	.84	1.13	1.20	1.00
	−	2	178	180					
		5	195	200					

tion of request terms, from information about each term considered independently. In general, it is assumed that terms are distributed independently and randomly. More specifically, we may adopt:

> *Independence Assumption* I1: The distribution of terms in relevant documents is independent and their distribution in all documents is independent;

or

> *Independence Assumption* I2: The distribution of terms in relevant documents is independent and their distribution in non-relevant documents is independent.

Assumption I1 is the basis for F1 and F3, while I2 underlies F2 and F4. It is unlikely in real life that index terms are assigned independently; but, in the absence of any more detailed information about co-occurrence probabilities, the independence assumptions form a reasonable starting point. It is argued in the Appendix that I2 is likely to be a better description of reality than I1; one of the objects of the experiments reported later is to establish whether this is the case and how much it matters.

In retrieval, some ordering principle is required for presenting output documents. In the absence of prior

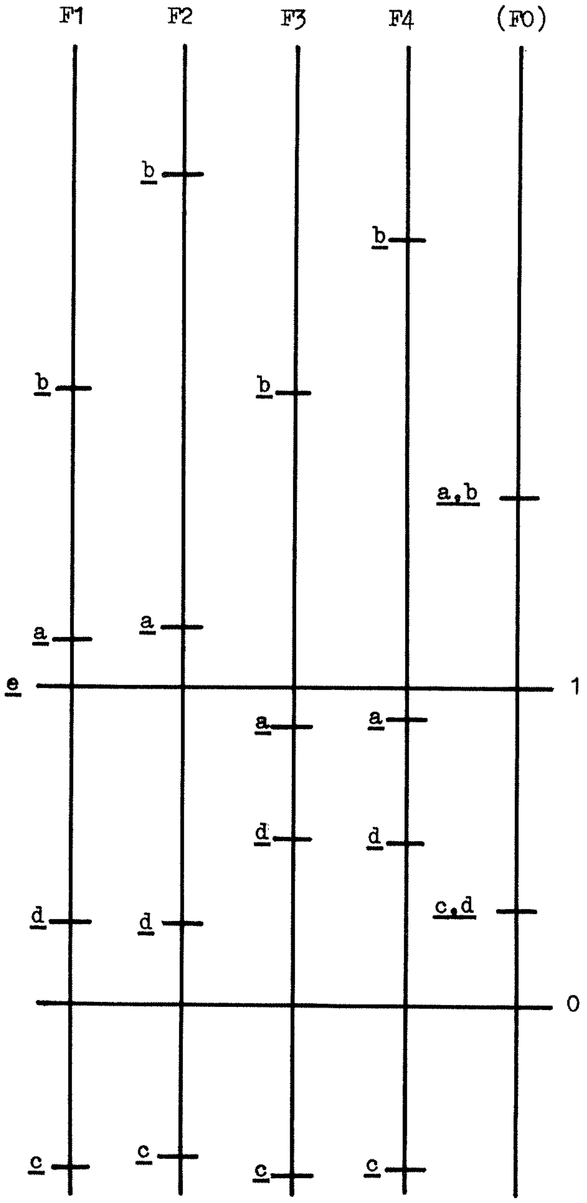

Fig. 1. Illustration of Comparative Weighting Values for Terms Using Different Formulae

knowledge, the only rational principle is that documents should be ordered by their probable relevance to the query. However, more specific principles are required to interpret probable relevance. The two possible principles are:

> *Ordering Principle* O1: That probable relevance is based only on the presence of search terms in documents;

and

> *Ordering Principle* O2: That probable relevance is based on both the presence of search terms in documents and their absence from documents.

In a sense, a dependence on term presence indicates an implicit dependence on term absence. However, the distinction between these principles concerns *explicit* recognition of term absence in the calculation of probabilities. Principle O1 in fact underlies F1 and F3, O2 F2 and F4. The first principle is more obvious and (since F1 is based on it) was implicitly assumed by Barkla and by Miller. But we argue in the Appendix that for optimal ranking, one should take explicit account of term absence: that is O1 is in some sense incorrect and O2 is correct. The argument is informally illustrated by the data given in Table 2 on the retrieval possibilities for two terms under Assumption I2, according to Functions F2 and F4 respectively. We retain Principle O1 and the functions derived from it to link our experiments with earlier work in the field.

To summarize: taking independence assumptions and ordering principles together, the theory yields four specific weighting functions as follows:

		Independence Assumptions	
		I1	I2
Ordering Principles	O1	F1	F2
	O2	F3	F4

Weighting Function F1 is based on the simplest and most obvious choices of Assumption and Principle, while F4 derives from more complex and less obvious ones. However, as argued above and in the Appendix, Ordering Principle O2 is correct and O1 incorrect, and Independence Assumption I2 is likely to describe reality more closely than I1. Thus, the theory predicts that F4 is the best of the four functions.

WEIGHTS AND DOCUMENT RANKING

Traditionally, the assignment of weights to index terms has been regarded as a separate issue from the formulation of a matching coefficient which can be used to rank retrieved documents. However, the theory specifies an explicit document ranking function: in order to derive a term weighting function, we have to assume that the matching coefficient consists of the sum of the weights of the matching terms. In other words, the theory specifies that these particular functions *should* be used with a sum of weights matching procedure. Other possible combinations of weighting function and matching coefficient would be compatible with the document ranking function: the most obvious example would be a non-logarithmic form of Functions F1-F4 coupled with

Table 2. Comparative effect of weights using weighting Functions F2 and F4.

Document Indexing		Document Relevance			F2	F4
		+	−			
Term *a*	+	5	20	25	.70	.95
	−	5	180	185		
		10	200	210		
Term *b*	+	8	50	58	.51	1.08
	−	2	150	152		
		10	200	210		

a and *b* share: $\frac{5 * 8}{10}$ = 4 relevant documents, and

$\frac{20 * 50}{200}$ = 5 non-relevant documents

The "richness" of, or proportion of relevant documents in,

Document set 1, indexed by both *a* and *b*, is $\frac{4}{9}$ = 44%;

Document set 2, indexed by *a* alone, is $\frac{1}{16}$ = 6%;

Document set 3, indexed by *b* alone, is $\frac{4}{49}$ = 8%.

Searching should retrieve document sets in order of richness. Both F2 and F4 retrieve set 1 first; F2 retrieves set 2 next, while F4 retrieves set 3. As set 3 is richer than set 2, F4 is preferable to F2.

a *product* of weights matching procedure. (Thus, it can be seen that the reason for using logs in all of the functions relates to the matching coefficient used.)

This slight latitude in the choice of weighting function and matching coefficient arises because any monotonic transformation of the document ranking function will produce the same ranking of the documents. But such monotonic transformations aside, the theory (together with the choice of Assumption and Principle) essentially determines both the weighting function and the matching coefficient.

In particular, it is not necessarily possible to derive from the theory a weighting function to go with *any* particular choice of matching coefficient. Consider, for example, Salton's cosine correlation. The main difference between cosine correlation and sum of weights as far as each individual question is concerned (and assuming binary index descriptions) is that cosine correlation is normalized for length of document, *i.e.*, number of index terms assigned to the document. Our theory, as presently developed, makes no predictions about the possible usefulness of document length as an indicator of relevance; it is indeed not clear whether in any more sophisticated theory, cosine correlation would be an appropriate way of using this information.

SOME PRACTICAL CONSIDERATIONS

Before applying the weighting functions we have described, several minor points need to be discussed. They are treated more adequately in the Appendix, in the context of the theory, and are simply summarized here to clarify the following account of our experiments.

(a) Estimation. The theory specifies the weighting functions in terms of probabilities: the probability of a document being posted to a term, given that it is relevant/non-relevant. We now have to *estimate*

the probabilities from the available information. The obvious estimate of a probability is a simple proportion; and such estimates were used for the versions of the four functions given previously. However, a proportion is not necessarily the best estimate for the purpose. We distinguish here between the two kinds of experiment we will be describing:

(i) If the weights are being used *retrospectively*, to determine optimal performance on a test collection, then the simple proportion estimates are the right ones to use; but

(ii) If the weights are being used *predictively*, in interactive retrieval, then other estimates are appropriate. We have used the simplest modified estimates, which are derived from the following modified contingency table:

		Document Relevance		
		+	-	
Document Indexing	+	$r + .5$	$n\text{-}r + .5$	$n + 1$
	-	$R\text{-}r + .5$	$N\text{-}n\text{-}R{+}r + .5$	$N\text{-}n + 1$
		$R + 1$	$N\text{-}R + 1$	$N + 2$

(b) *Scaling.* The prime output of a search exploiting a relevance weighting formula is an ordered set of documents. In operational situations, a cutoff by number of documents would presumably be applied. Individual values of the matching coefficient are in principle of no utility. However, for test purposes we need to average over a set of requests, which implies that the matching values should be comparable between different requests. As they stand, Weighting Functions F1 and F2 satisfy this criterion, but F3 and F4 do not. In the experiments we have adopted one of the two stratagems suggested in the Appendix: that of computing a "correction value" to be added to the matching coefficients of all documents in relation to a particular request.

(c) *Limiting Cases.* In the retrospective case, where the simple formulae are used, it will be evident that problems arise whenever any of the components of the formulae are zero. Some of these are, so to speak, external in that matching is impossible, or necessarily undiscriminating. This applies where N, R, $N\text{-}R$, n or $N\text{-}n = 0$. The obvious weight for a term so affected is 0. The more important limiting cases, and their resolution, are indicated in Table 3. (Some suitable coding of these cases is of course needed for retrieval programs.)

• Experiments

In previous experiments (*10*), Weighting Function F1 was compared with simple collection frequency weighting F0, and with unweighted term matching. As mentioned, relevance weighting may be applied in two different ways: using perfect information to give optimal performance or predictively. Optimal performance for F1 was enormously better than that for unweighted terms or terms weighted by F0, for all three collections used. Predictive performance, using F1 without the modification to the estimates mentioned above, was not really superior to that for weighting by F0, but was materially better than that for simple term matching. An extended range of comparisons for F1 to F4, for a larger test collection, is described here.

The tests were carried out with the manually indexed Cranfield 1400 collection. Collection details and results are summarized in Table 4. Matching involves either real or notional coordination levels; averaging over the set of requests is by simple summing of document numbers retrieved; and performance is represented by recall-precision graphs with precision values obtained for standard recall levels by interpolation. Fig. 2 illustrates performance for simple term matching and weighting using F0, and F1 to F4. All the weighting schemes are optimal in that the distribution of terms in documents and relevant documents is known. Fig. 3 shows a predictive test (with modified estimates) of the relevance weighting formulae, with performance compared to simple term matching and also collection frequency weighting, the latter of course without any predictive element. This experiment was conducted by dividing the complete set of documents into its equal odd and even-numbered subsets and applying weights calculated from the even-numbered subset to the odd-numbered one. In contrast, Fig. 4 shows relevance weighting based on the odd-numbered set itself: *i.e.,* Fig. 4 gives the optimal performance for this subset.

These results show that the performance level of optimal relevance weighting is very much higher than that for simple terms or terms weighted by collection frequency*. Predictive relevance weighting as illustrated in Fig. 3 appears much less useful, although still

*A rough and ready significance test requires a 5 percent difference in the areas enclosed by two curves on such a graph. All differences commented on in the text are significant in this sense.

Table 3. Treatment of special cases for weighting functions.

Case	Definition	Documents in Which Term Occurs	Functions to Which Case Applies	Implications of Case for Document When Term Is:	
				Present	Absent
A	$r = 0$	Non-relevant only	F1,F2,F3,F4	Bad	Indifferent
B	$n\text{-}r = 0$	Relevant only	F2,F4	Good	Indifferent
C	$R\text{-}r = 0$	All relevant and some others	F3,F4	Indifferent	Bad
D	$N\text{-}n\text{-}R+r = 0$	Some relevant and all others	F4	Indifferent	Good
E	$n\text{-}r = 0$ $R\text{-}r = 0$	All relevant and no others	F4	Good	Bad
F	$r = 0$ $N\text{-}n\text{-}R+r = 0$	No relevant and all others	F4	Bad	Good

"Bad" means that the document should never be retrieved, *i.e.* should be at bottom rank;

"Good" means that the document should always be retrieved, *i.e.* should be at top rank;

"Indifferent" means that the document should be unaffected, *i.e.* should be at the rank determined by its other terms.

Case E combines B and C; case F, A and D.

Cases C through F apply to functions F3 and F4, in which term absence is explicitly recognized.

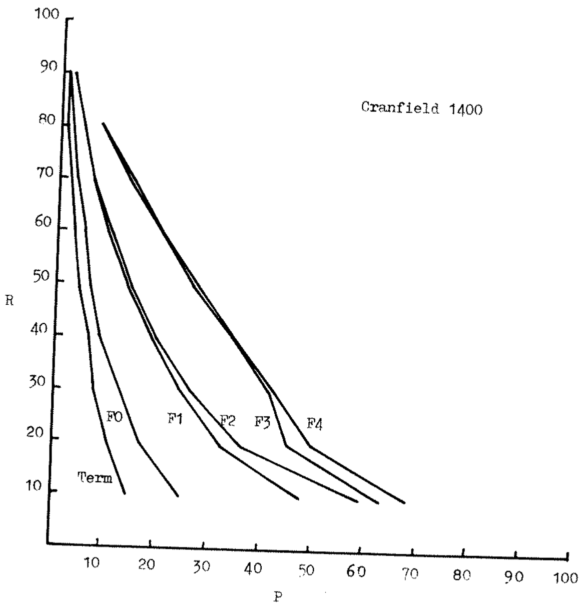

Fig. 2. Comparative Retrieval Performance Using Weights for Cranfield 1400 Collection

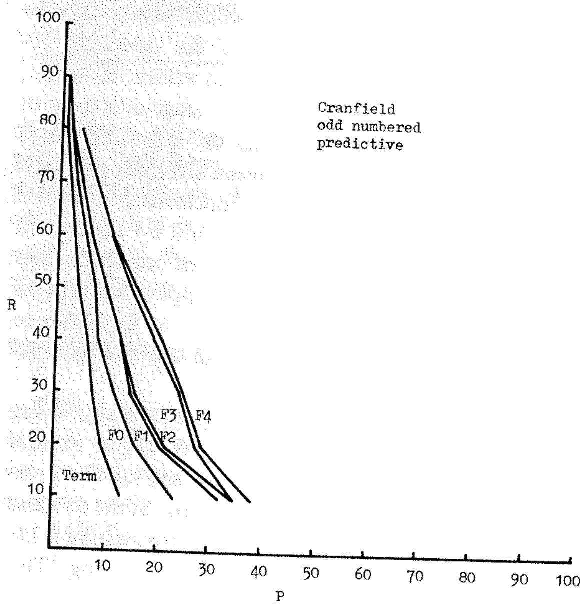

Fig. 3. Comparative Retrieval Performance Using Weights Predictively for Cranfield 700 Odd-Numbered Sub-collection

Table 4. Test collection properties and retrieval results.

	Cranfield 1400	Keen
Number of documents	1400	797
Number of requests	225	63
Number of relevant documents	1614	936
Number of terms	2683	939
Average terms per document	29.9	7.2
Average terms per request	7.9	5.3
Average relevant documents per request	7.2	14.9

Precision

	Cranfield 1400						Cranfield Odd Numbered					
Recall	Terms	F0	F1	F2	F3	F4	Terms	F0	F1	F2	F3	F4
100	–	–	–	–	–	–	–	–	–	–	–	–
90	1	1	2	2	–	–	1	1	2	2	7	7
80	1	2	4	4	8	8	1	2	4	4	13	12
70	2	3	6	6	13	14	2	3	7	7	21	20
60	3	5	9	10	19	19	3	5	11	12	27	29
50	4	6	13	14	26	27	4	7	16	17	35	35
40	6	8	18	19	34	34	6	8	21	23	41	43
30	7	12	24	26	41	42	7	11	29	31	48	50
20	10	16	32	36	45	49	9	15	37	42	56	57
10	14	24	48	59	63	68	13	23	52	69	67	78
	Cranfield Odd Numbered Predictive						Keen					
	Terms	F0	F1	F2	F3	F4	Terms	F0	F1	F2	F3	F4
100	–	–	–	–	–	–	–	–	–	–	–	–
90	1	1	1	1	–	–	–	–	–	–	–	–
80	1	2	2	2	4	4	–	–	–	–	–	–
70	2	3	4	4	7	7	6	6	8	8	10	10
60	3	5	6	6	10	10	6	7	10	10	17	17
50	4	7	9	9	14	15	7	11	12	15	23	24
40	6	8	12	12	19	20	10	16	22	25	31	34
30	7	11	14	15	24	24	15	22	34	41	45	46
20	9	15	20	21	27	28	20	37	46	52	57	61
10	13	23	32	35	35	38	37	51	70	73	69	77

materially better than either unweighted terms or collection frequency weighting. Functions F1 and F2, and F3 and F4 respectively perform the same, but the last two are very much superior to the first two.

For comparison, results for retrospective weighting for another collection, Keen, are illustrated in Fig. 5. Relative performance for the different weighting schemes is the same as for the Cranfield collection.

• Discussion of the Results

Two immediate conclusions can be drawn from the results of our experiments. First, our argument that Ordering Principle O2 is correct and O1 incorrect is confirmed by the experiments; F3 and F4 performed consistently better than F1 and F2. Secondly, our assumptions of term independence seem not to be critical ones, since the choice of I1 or I2 had virtually no influence on performance. However, it could be that with larger and/or more heterogeneous collections the choice would affect performance.

The large difference between the retrospective and the predictive uses of the weighting functions indicates the importance of the estimation problem. In the predictive situation, we are trying to estimate the values of the probabilities defined in the Appendix from fairly small samples of documents (particularly of relevant documents); it is therefore not surprising that performance is

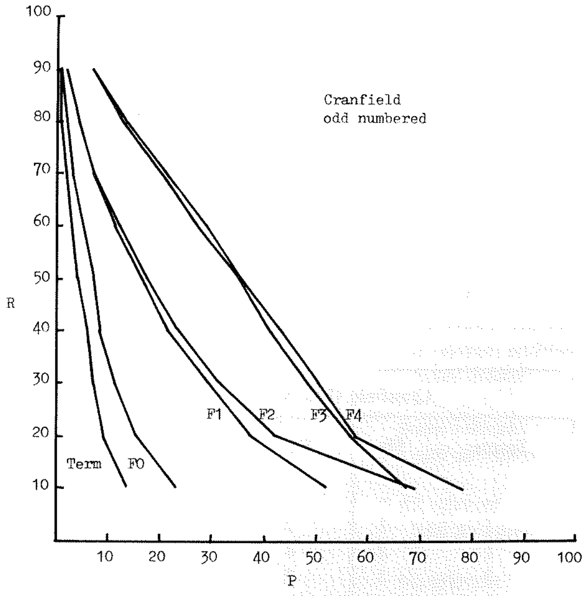

Fig. 4. Comparative Retrieval Performance Using Weights for Cranfield 700 Odd-numbered Sub-collection

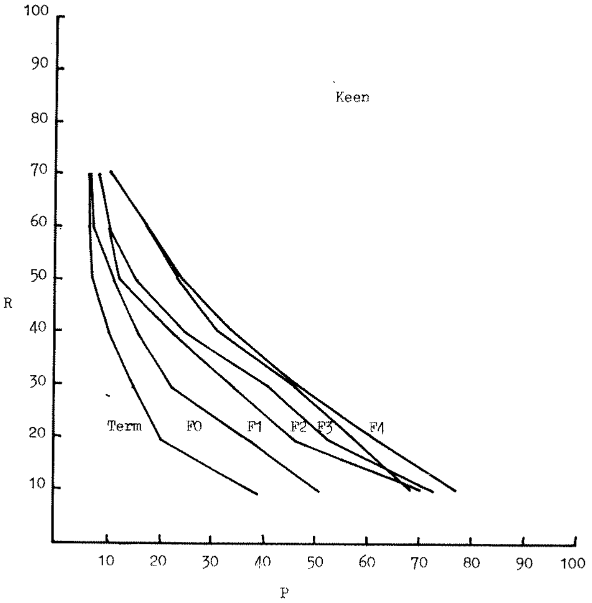

Fig. 5. Comparative Retrieval Performance Using Weights for Keen 797 Collection

not so good as in the retrospective case. Indeed, the small sample size may have the effect of *improving* performance in the retrospective case; the weighting functions in effect make use of *any* statistical properties of the particular test collection, even though they may be random properties of the sample rather than meaningful properties of the terms and questions. If there is only one relevant document, for example, any of the four functions can be almost guaranteed to find a term combination that retrieves that document and practically no others.

The problem of estimation deserves further study; some ideas on the subject are discussed in the Appendix. The general problem is: what information can we use, and how can we best use it to estimate the value of F4 (or whatever weighting function we choose)? Apart from relevance feedback data, which is all we have used so far, it may be helpful to make use of the questioner's knowledge of the terms (as Miller does), and/or the results of previous tests on the system. But, in order to make use of such information, we need to extend the theory to include it. Further, we need to know how to use partial relevance feedback data, such as that which is provided when the *output* of a previous search is judged for relevance, but not the whole collection. A possible method of doing this is described in the Appendix.

It should, however, be stressed that our experimental results indicate that even very crude estimation techniques can provide improvements in performance over searching without relevance weighting.

• Other Experiments

Some comments on the relation between our results and those obtained elsewhere are required. Two series of tests in particular are comparable in scale and intention.

Miller's experiments (*8, 9*), as mentioned earlier, were based on Weighting Formula F1. This was applied predictively, incorporating (supposed) user estimates of the frequency of applicability of terms to relevant references. A MEDLARS data base containing 210,000 documents was used, with 25 search formulations, the main test being to compare performance for the weighted probabilistic search with that of standard Boolean searching, on MeSH indexing. Output was cutoff for the former to give approximately the number of documents retrieved by the latter. Relative recall can be obtained as the percentage of relevant documents obtained by either method, *i.e.*, retrieved by each.

The results show precision values of 17 and 15.5 percent, and recall values of 69 and 95 percent, for Boolean and probabilistic searching respectively. Thus, probabilistic searching leads to a striking improvement in recall, with no real loss of precision. It is unfortunate that so few requests were involved, and that only limited performance comparisons were made. Further tests applied probabilistic weighting to title words. This also gives slightly better recall than Boolean searching on index terms, but precision is somewhat reduced. The difficulties of drawing general inferences from Miller's experiments are illustrated by the fact that, if relative recall is based on the pooled relevant documents retrieved by all three strategies, values are much lower than in the pairwise comparisons. On the other hand, probabilistic searching still comes out well, specifically with 64 percent recall on indexing, compared with 46.5 percent for Boolean search. It should be noted that these experiments suggest that it is unlikely that the choice of document cutoff for probabilistic searching would present problems: an arbitrary low one would seem quite effective.

Miller's investigations thus confirm the value of relevance weighting, in a rather different framework.

The second set of tests has been carried out by the SMART Project. As noted above, early SMART experiments (*12*) with relevance weighting, used retrospectively, were trivial in scale. More ambitious tests applying the same formula predictively to two collections of 155 requests and 424 documents, and 83 requests and 425 documents, are described in a recent report (*13*). The results show no useful improvement in performance, which would seem to cast doubt on our findings. However, closer examination shows that the experiments constituted a rather eccentric and very partial test of relevance weighting. In particular, they involved a very drastic form of prediction.

Instead of splitting the document set as we did, the SMART workers split the request set, using weights obtained from one to control searches with the other. Further, any term occurring in more than one request in the first set was assigned the average of its relevance weights. One important reason for the small overall effect of the weights must be that relatively few terms occurred in queries in both sets, and so could be assigned weights. Also, only the ten highest weighted terms in a query were selected for weighting. The net effect of these two restrictions was that only about 10 percent of the request terms in the search set were assigned relevance weights at all. (The effect of the weights was additionally obscured by the use of within-document term frequencies, and of term phrases and classes as well as single terms.)

There is, however, a more fundamental problem about the assumption that the use of a term in one question tells us something about its use in another. We believe that SMART workers need to assume that if term t occurs in questions A and B, the probability that t is assigned to a document relevant to A is the same as the probability that it is assigned to a document relevant to B. There is no evidence for such a bold assumption, and it indeed seems somewhat unlikely to hold.

The SMART approach differs from ours in the treatment of both documents and requests. An index term may be weighted with respect to a document set or individual documents, or with respect to a request set or individual requests. In the range of weighting formulae we have considered, an individual term has the same weight in different documents. Salton normally allows terms different weights in documents, depending on their within-document frequency, and utilizes collection-based weights as well. On the other hand, Salton starts with the same precision value for a term in different requests (though its final weights differ due to other factors), where we have allowed different values, *i.e.* weights. Further study of the most effective ways of exploiting the four sources of information about terms is clearly desirable.

• Conclusions

In this paper (with its Appendix), we present a theory of search term weighting exploiting relevance information and suggest that a particular formula assuming distinct term distributions in relevant and non-relevant documents, and ordering documents in matching by both term presence and term absence, exploits this information most effectively. Our experiments show that retrieval performance improves when information about the occurrences of terms in relevant documents is added to information about their simple document incidence; and these experiments confirm the superiority of the preferred formula.

More specifically, relevance weights, whether retrospective or predictive, give noticeably better performance than simple term matching. When relevance weights are applied retrospectively (*i.e.*, are optimal) a strikingly high level of performance is achieved, which can, as Sparck Jones suggested earlier (*10*), be taken as a general experimental yardstick. Predictive relevance weights are also effective. When used as originally suggested for SDI or iterative searching, performance may be expected to improve with cumulating information, hopefully to converge with the ideal case.

It has been suggested that this type of statistical weighting scheme is simply a method of improving performance suited to largely automatic systems with little user intervention; and that it is unlikely to give better performance than retrieval based on careful user request formulation, particularly when supported by iterative searching of partial or full document files. None of the experiments to date throw any light on this question, since only raw requests were input in our tests and in the SMART ones while the request terms in Miller's were taken from the prior carefully formulated Boolean search statements; no iterative experiments have been undertaken. Further work is clearly needed to identify especially convenient and effective combinations of system information and user effort.

It should be noted that term values based on relevance information may be used for purposes other than to control matching as described here. Barker, Veal and Wyatt (*11*) used them to select terms from retrieved documents to amplify requests. This use is related to other request modification schemes like those studied under the heading of relevance feedback by the SMART Project (*14*). This is another topic inviting further research. It is also worth noting that the theory as presented in this paper does not concern traditional index terms only: in principle, it applies to *any* key that might usefully be taken as an indicator of relevance (*e.g.*, authors, citations, journals, etc.) as well as, of course, any kind of retrieval language, natural or controlled. There may be more problems with the independence assumptions in some specific areas (*e.g.,* precoordinate systems); again, further investigation is required.

There are indeed many possibilities and questions in this area but we believe that the theory and experiments we have presented demonstrate the value of a systematic statistical use of relevance information.

Acknowledgments

We are very grateful to Dr. C.J. van Rijsbergen for his comments on the manuscript of this paper, and to our referees.

Appendix: A Probabilistic Theory of Relevance Weighting

We start from the intuitively obvious proposition (the "probability ordering principle") that the best rank ordering of a set of documents for presentation to a user is that in which the documents most likely to be relevant to his request are nearest to the top. We are then faced with the question: what information can we use and how can we use it to assess the probability of relevance of any given document?

INDEPENDENCE ASSUMPTIONS

We assume in this paper that the available information consists of term distributional information; information concerning the frequencies of occurrence of terms in relevant and other documents, given binary document descriptions. We assume that we do *not* have term co-occurrence information; to make up for this lack, we have to make some assumptions about term co-occurrences. Specifically, we assume that the terms occur independently; more specifically, we have two alternative sets of assumptions:

I1a. The occurrences of different terms are independent within the set of relevant documents.

I1b. The occurrences of different terms are independent within the whole collection.

I2a. The occurrences of different terms are independent within the set of relevant documents.

I2b. The occurrences of different terms are independent within the set of non-relevant documents.

Miller (*8, 9*) and Barkla (*7*) both make Assumptions I1a and b; Yu and Salton (*12*) make Assumptions I2a and b.

Robertson (*6*) points out that Assumptions I1a and I1b are not strictly compatible; since request terms normally occur more frequently in relevant than in non-relevant documents, I1a suggests (and in some cases implies) that terms from the same question must co-occur more frequently in the whole collection than would be expected under I1b. Indeed, this appears to be the case. Since Assumptions I2 would predict this result anyway, we prefer those assumptions.

The use of any independence assumptions at all is suspect, since they certainly do not hold universally. The alternative would be to look for term co-occurrence information: Barker, Veal and Wyatt (*11*) consider using term pairs. But very much more data would be required to make adequate estimates of the parameters necessary to define co-occurrence properties; and, as the discussion of our results indicates, we already have problems in finding out enough about the terms alone from the usual test collections. In any case, our results indicate that the independence assumptions are not really critical, performance for different assumptions (with one ordering principle) being the same.

ORDERING PRINCIPLES

Although the probability ordering principle itself is intuitively obvious, its application in a particular situation is not so obvious. Barkla and Miller both assume implicitly that the probability of relevance of a particular document should be calculated simply on the basis of the terms in that document; that is, they ignore any request terms missing from the document. But, potentially at least, the absence of a term from a document carries some information about that document and, if this information has any bearing on the probability of relevance, then we should make use of it. In the text we gave an example to show the difference between the two approaches; the example shows that we should take absence into account. Robertson (*15*) presents a more formal proof that the probability ordering principle, with the probabilities calculated from *all* available information, leads to the best possible expected performance (in a sense to be discussed later on in this article).

We therefore consider two ordering principles as the basis for the formal analysis which follows:

O1 The probability of relevance of a document should be calculated from the terms present in the document only.

O2 The probability of relevance of a document should be calculated from the terms present in the document and from those absent.

These two principles are not alternative or potentially equally valid assumptions about the world like I1 and I2; on the contrary, theoretically, O2 is correct and O1 is incorrect. But we include O1 as a possible ordering principle in order to relate our work to earlier work, hoping to confirm empirically our theoretical demonstration of the superiority of O2.

In the following three sections, we use Ordering Principle O2 and Independence Assumptions I2 to develop the Weighting Function F4. The other three functions described in the text are derived in exactly parallel ways from different combinations of the principles and assumptions; the differences of detail are described below.

THE PROBABILITY OF RELEVANCE

We consider a request consisting of a set of (initially unweighted) terms, which we denote by Q. Each request term t_i has two probabilities associated with it:

$$\psi_{i1} = P\,(\text{document contains } t_i \mid \text{document relevant})$$

(that is: the probability that a document contains the term t_i, given that it is relevant), and

$$\psi_{i2} = P\,(\text{document contains } t_i \mid \text{document non-relevant}).$$

"Relevant" is taken to mean relevant to the need underlying the request: thus, the probabilities are need specific. We assume that the available relevance information allows us to estimate these probabilities; the question of estimation is discussed below.

We consider subsets T of the set of request terms Q, *i.e.* $T \subseteq Q$. We define D_T as the set of documents matching the request on *exactly* the terms T, implying that they contain none of the terms in Q - T, and

$$\phi_T = P\,(\text{document relevant} \mid \text{document is in } D_T)$$

(any document is in exactly one set D_T). Finally,

$$\phi_C = P\,(\text{document relevant} \mid \text{document is in collection})$$

(in the notation of the text, ϕ_C is estimated by R/N).

In what follows, we make considerable use of the well-known logistic (or log-odds) transformation of a probability:

$$\text{logit}\, P = \log \frac{P}{1\text{-}P}\,.$$

This transformation is strictly monotonic. We make use of it simply to put a complex function in a more convenient linear form.

We now want to express the probability of relevance of any document ϕ_T in terms of the probabilities ψ_{i1} and ψ_{i2}. This is accomplished as follows.

By successive application of Bayes' Theorem, it can be shown that, for any two events a, b,

$$P(a/b) = \frac{P(b/a)\,P(a)}{P(b/a)\,P(a) + P(b/\bar{a})\,P(\bar{a})}$$

($\bar{a}$ is "not a"). It follows that

$$\text{logit}\, P\,(a/b) = \log \frac{P(b/a)}{P(b/\bar{a})} + \text{logit}\, P\,(a) \qquad (1)$$

If a is "document relevant" and b is "document in D_T", then

$$P\,(a/b) = \phi_T \text{ and } P\,(a) = \phi_C.$$

From Independence Assumption I2a, we deduce that

$$P\,(b/\bar{a}) = \prod_{t_i \in T} \psi_{i1} \prod_{t_j \in Q\text{-}T} (1\text{-}\psi_{j1})$$

and from I2b,

$$P\,(b/a) = \prod_{t_i \in T} \psi_{i2} \prod_{t_j \in Q\text{-}T} (1\text{-}\psi_{j2}).$$

Now Equation (1) becomes

$$\text{logit}\, \phi_T = \sum_{t_i \in T} \log \frac{\psi_{i1}}{\psi_{i2}} + \sum_{t_j \in Q\text{-}T} \log \frac{(1\text{-}\psi_{j1})}{(1\text{-}\psi_{j2})} + \text{logit}\, \phi_C. \qquad (2)$$

We can use this equation to rank the documents in order of their probability of relevance ϕ_T. The reason for the use of the logit function is that the equation is then in a linear form, which is convenient for use in a weighting scheme, as we shall see below. Since logit is a monotonic function, logit ϕ_T ranks the documents in the same order as ϕ_T.

RETRIEVAL AS A DECISION PROBLEM

In the context of Bayesian decision theory, Equation (2) could be used as the basis for a discriminant function, as follows. We define a "loss function," associated with the decision as to whether or not to retrieve a document:

$$\text{Loss (retrieved} \mid \text{not relevant)} = a_1$$

(that is, the loss associated with retrieving a non-relevant document is a_1);

$$\text{Loss (not retrieved} \mid \text{relevant)} = a_2 .$$

A document in D_T has a probability ϕ_T of being relevant. So if we retrieve it, the expected loss will be

$$(1\text{-}\phi_T)a_1 .$$

If we do not retrieve it, the expected loss will be

$$\phi_T a_2 .$$

So the optimum (loss-minimizing) decision rule is to retrieve if

$$\phi_T a_2 > (1\text{-}\phi_T)\, a_1$$

or $$\frac{\phi_T}{1\text{-}\phi_T} > \frac{a_1}{a_2}$$

or $$\text{logit } \phi_T > \log \frac{a_1}{a_2}$$

Thus, if we define

$$g_T = \text{logit } \phi_T - \log \frac{a_1}{a_2} ,$$

then g_T is an optimum linear discriminant function [in the terms of Nilsson (*16*)], with the corresponding decision rule:

$$\text{Retrieve if } g_T > 0. \qquad (3)$$

Our approach, however, is to simply rank the documents and allow a cutoff decision to be made by the user. If we rank the documents by logit ϕ_T, then a decision rule equivalent to Equation (3) could be applied by the user:

$$\text{Retrieve if logit } \phi_T > \log \frac{a_1}{a_2} .$$

But other forms of retrieval rule may be appropriate [Cooper (*17, 18*) has reviewed a number of different kinds]. It must be stressed that the ordering principle is independent of the particular decision rule chosen, whether or not the decision rule can be represented by a simple loss function such as that given previously. The *only* assumption we make is that users prefer relevant to non-relevant documents.

In an operational system, with a large document collection, it may be quite impracticable to rank the whole document collection; and a cutoff point must therefore be specified, above which retrieved documents only are ranked. An approach compatible with the ideas presented in this paper would be to set the cutoff lower than any users are likely to set it. An alternative would be a decision theoretic approach such as that indicated above, where a_1 and a_2 are values which have to be provided by the user in some form.

A very much more complex decision-theoretic retrieval model is presented by Tague (*19*).

THE WEIGHTING FUNCTION

We could consider implementing a retrieval system which made direct use of Equation (2) in order to rank the documents. However, we prefer to achieve the same effect by means of a simple term weighting scheme for the following reasons:

(a) Most previous work in this area has used term weighting schemes.

(b) Term weighting is well understood and simple to implement; indeed many systems exist which can perform weighted term searches.

It should be pointed out that a theory which produces a ranking equation such as (2) cannot *necessarily* be translated into a term weighting scheme. The necessary condition for such translation is that it should be possible to put the ranking equation into a linear form. We have already done so with Equation (2) by making use of the logit transformation.

We now want to use Equation (2) to derive weights for the request terms, which will have the effect of ranking the documents in the required order. We have two possible courses of action; the first is as follows:

We assign to each term t_i a weight

$$v_i = \log \frac{\psi_{i1}}{\psi_{i2}} \qquad (4)$$

to be given to any document that contains the term; we *also* assign a weight

$$u_i = \log \frac{(1\text{-}\psi_{i1})}{(1\text{-}\psi_{i2})} \qquad (5)$$

to be given to any document that does *not* contain the term. Because this scheme requires one to take account of the absence of a term as well as its presence, we will call it a P/A-weighting scheme. Then a document in D_T gains a total weight (matching value) of

$$\sum_{t_i \in T} v_i + \sum_{t_j \in Q-T} u_j$$

which by equation (2) is equal to

$$\operatorname{logit} \phi_T - \operatorname{logit} \phi_C \ .$$

Hence, this weighting scheme does rank the documents in order of ϕ_T. Furthermore, this matching value has some significance other than just as an ordering mechanism; this can best be seen if we consider a document with matching value zero, which would imply

$$\operatorname{logit} \phi_T = \operatorname{logit} \phi_C, \text{ or } \phi_T = \phi_C.$$

In other words, a document with matching value zero has the same probability of relevance as a document chosen at random from the collection.

A simple weighting scheme in the usual sense takes account of the presence of a term only; we will call this a P-weighting scheme. We can devise a P-weighting scheme equivalent to the above P/A-weighting scheme as follows. We assign to each term t_i a single weight

$$w_i = v_i - u_i = \log \frac{\psi_{i1}(1-\psi_{i2})}{\psi_{i2}(1-\psi_{i1})} \qquad (6)$$

which is given to a document that contains the term. Then a document in D_T has a matching value of

$$\sum_{t_i \in T} v_i - \sum_{t_j \in T} u_j$$

which by Equation (2) is equal to

$$\operatorname{logit} \phi_T - \operatorname{logit} \phi_C - \sum_{t_j \in Q} u_j \ . \qquad (7)$$

This matching value still ranks documents in ϕ_T-order, since the last sum of Equation (7) does not depend on the particular set of matching terms T, but rather on all the terms in the request. However, we have now lost the significance of the matching value itself; the only significance is in the ordering it gives to the documents.

If we replace the probabilities in Equation (6) by the equivalent proportions, ignoring the subscript i and using the notation defined in the text of the paper, we find

$$w = \log \frac{\psi_1}{1-\psi_1} \bigg/ \frac{\psi_2}{1-\psi_2}$$

$$= \log \frac{r}{R-r} \bigg/ \frac{n-r}{N-n-R+r} \qquad (8)$$

which is the Weighting Function F4 as defined in the text. However, before doing this, we need to consider carefully the problem of *estimating* the probabilities from the available data.

(The derivations of F1 to F3 follow exactly parallel lines; the differences are as follows. Using O1 rather than O2, for F1 and F2, Equation (2) would lose the part relating to Q-T; thus we would not have the problem later on of getting rid of the term-absence weights. Using I1 rather than I2, for F1 and F3, we would replace ψ_{i2} by the equivalent probability for *all* documents; then instead of using the logistic transform, we would use a simple log.)

ESTIMATION

How do we estimate the probabilities ψ_{i1} and ψ_{i2}? In order to answer this question, we must distinguish (as we have done in the text) between the two possible uses of the weighting schemes. The first is to use the schemes retrospectively on a test collection to give optimal performance. In this case we have perfect information about the term distributions; if there are R relevant documents and the term occurs in r of them, then we *know* that ψ_{i1} (for this collection) is r/R. Thus, there is no problem about estimation and we can use the formulae given in the text without alteration.

The second possibility is to use the weighting scheme predictively, using relevance data from one search to improve the search strategy for the next. In this case, we are trying to make inferences about the probabilities on the basis of sample information. Although the simple proportions used in the retrospective case make reasonable estimates of the probabilities for many purposes, there are various reasons against their use in our situation. The main problem is that the samples are often small (the number of relevant documents in particular), coupled with the fact that we are trying to estimate not the probabilities themselves, but non-linear functions of them.

The Weighting Function (6) arises fairly commonly in a variety of contexts [see Cox (*20*)] and there has been a fair amount of work on methods of estimating it. Cox suggests a straightforward modification of the simple proportion method of Equation (8), which involves adding ½ to each of the four quantities in the expression:

$$w = \log \frac{r+\frac{1}{2}}{R-r+\frac{1}{2}} \bigg/ \frac{n-r+\frac{1}{2}}{N-n-R+r+\frac{1}{2}} \ . \qquad (9)$$

(This procedure may seem somewhat arbitrary, but it does in fact have some statistical justification.) This is the method we have adopted when using the schemes predictively; its application to the other three functions is defined by the following adjusted 2 X 2 table:

$r + ½$	$n\text{-}r + ½$	$n + 1$
$R\text{-}r + ½$	$N\text{-}n\text{-}R+r + ½$	$N\text{-}n + 1$
$R + 1$	$N\text{-}R + 1$	$N + 2$

Although Cox's justification for this procedure applies only to F4 directly, we have used the method for F1 to F3 on the grounds that it is likely to provide better estimates than the simple proportion method, and a statistically watertight method would probably be much more complex.

We could, indeed, take a much more sophisticated approach to the whole estimation problem. Robertson and Teather (*21*) develop an estimation method for information retrieval test data which incorporates and provides estimates of Weighting Function (6). The method is Bayesian and makes use of the results for all the terms of the test questions in order to improve the estimates for each one. They find that under a fairly simple model of the relationship between different terms, the estimates of Weighting Function (6) for each term tend to converge, confirming the point made earlier that we normally have rather little information on which to base the individual estimates. With a more general model, Robertson (*15*) shows that the major variation in the value of Function (6) appears to be related to variations in the specificity of terms; more specific terms (*i.e.* less frequent in the collection as a whole) tend to have higher values of Function (6). This confirms Sparck Jones' (*3*) results on the value of simple collection frequency weighting.

More work is required in this area. What is needed is a general estimation method which will make the best use of all available information to estimate Weighting Function (6) (or any modified version that is found useful). "All available information" might include:

- prior information, from formal tests, on the usual behavior of terms in the system;
- the questioner's prior expectations of, or knowledge of, the terms he wants to use;
- the frequencies of the terms he wants to use in the collection as a whole;
- any available relevance feedback data.

One particular point that may cause problems is the form of the relevance feedback data. We have assumed throughout this paper that it is complete, in the sense of relating to a completely evaluated collection, in which all relevant documents are known. Normally, however, if the system were being used in a practical situation (iteratively or for SDI), only the *output* of a previous search would be judged for relevance. How can we make use of such partial information?

The independence assumptions indicate a possible answer to this question. We consider Assumptions I2, and the probabilities ψ_{i1} and ψ_{i2} defined earlier. Under Assumption I2a, the probability ψ_{i1}, which is defined as

P (document contains t_i | document relevant),

is also equal to

P (document contains t_i | document relevant and matches search statement A_i),

where A_i is *any* search statement not containing t_i.

The other probability ψ_{i2} relates in the same way to the non-relevant documents. So in principle, we can estimate the probabilities, if we have a suitable search statement A_i, not containing term t_i, such that all the documents which match it have been evaluated.

Suppose for example, that on our first run we give all terms equal weight (one), and retrieve against a threshold of two. For a particular term t_i, we have to remove t_i from the search statement, but keep the threshold at two. Thus, the corpus we use to estimate ψ_{i1} and ψ_{i2} consists of those documents retrieved in the first search, which are presumably all judged for relevance, *less* those which would not have been retrieved had t_i not been in the original statement. This corpus should (under Assumption I2) provide unbiased estimates of probabilities.

It should be pointed out that the obvious method of estimation, using the output of the first search as it stands (as done by Barkla for example), is likely to give biased results according to the above analysis.

SCALING

The derivation of the weighting function assumed that the only important result of weighting was the final rank ordering of the documents and that the actual matching values were not significant. As we discovered in the derivation, the use of a P/A-weighting scheme gave matching values which were significant, but when we changed to the simpler P-weighting scheme, this significance was lost. (This problem applies to F3 and F4, but not to F1 or F2.)

There are some circumstances in which significant matching values would be desirable. In particular, some users might prefer a matching value which had some direct bearing on the probability of relevance of a specific document, rather than simply indicating its relation to other documents. Also, when testing such a system, one might want values which are comparable between questions for averaging purposes. (In fact, we require comparable values for our tests, as indicated in the text.)

The obvious way to achieve this is to reformulate F3 and F4 in a P/A-weighting form. The formulae, using simple proportion estimates, are as follows:

	v (term presence weight)	u (term absence weight)
F3′	$\log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n}{N}\right)}$	$\log \frac{\left(\frac{R-r}{R}\right)}{\left(\frac{N-n}{N}\right)}$
F4′	$\log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n-r}{N-R}\right)}$	$\log \frac{\left(\frac{R-r}{R}\right)}{\left(\frac{N-n-R+r}{N-R}\right)}$

Alternatively, we could keep the P-weighting forms F3 and F4, but add a constant to the matching values of all documents for the given question to restore the scale properties of the matching value. The appropriate constant is given by the rogue sum on the right hand side of Equation (7); it is in fact the sum of the u-values for all the terms in the question.

LIMITING CASES

As noted in the text, the use of simple proportion estimates can lead to problems if any of the quantities used are zero. Considering only the "internal" cases (see text), situations might arise in which one of the weighting functions yields:

$$\log 0 = -\infty,$$

or

$$\log x/0 = +\infty \quad .$$

These infinite weights should be interpreted in the obvious way; a document given a weight of + ∞ should be retrieved at the highest possible level, since it is certainly relevant and one with - ∞ should never be retrieved since it is certainly not relevant.

The exact application of these rules is obvious for F1 and F2, but for the other two functions we need to examine the P/A-weighting forms F3′ and F4′, since either extreme value may apply either to term presence or to term absence. The various possible cases are given in Table 3 in the text.

EXPERIMENTAL CONFIRMATION OF THE THEORY

On a superficial level, we can take the theory as a reasonable justification for trying out the various weighting functions in an experimental situation. More particularly, we can consider testing the weighting functions against unweighted retrieval using a standard testing methodology. However, if we wish to consider such experiments as direct tests of the theory itself, then we need to examine more closely the relationship between the probability ordering principle and the measures of performance used in such a test.

As indicated previously, Robertson (*15*) provides a formal proof that the probability ordering principle optimizes performance. In this proof, performance is measured in terms of a curve relating two probabilities:

$$\theta_1 = P(\text{document retrieved} \mid \text{document relevant})$$
$$\theta_2 = P(\text{document retrieved} \mid \text{document non-relevant}).$$

These probabilities are estimated by the obvious proportion measures, *recall* and *fallout;* or to put it the other way round:

θ_1 = expected recall (using "expected" in the statistical sense)
θ_2 = expected fallout.

Since recall, fallout and precision are related in a straightforward way, a recall method which can be expected to optimize recall/fallout can also be expected to optimize an experimental recall/precision curve.

The proof as it stands relates only to a single question. But because of the small numbers of documents involved, we have to average over a number of questions in order to get reasonable estimates of recall and precision. Thus, the standard methodology of retrieval experiments, which involves comparing average recall and precision curves, would seem to be a reasonable test of the theory.

This summary discussion ignores a number of statistical problems, such as the validity of averaging over questions, which are discussed in more detail elsewhere (*21, 15*). It is clear that a more rigorous test of the theory could be designed, although the methodology for such a test is not obviously available at present. In the meantime, the fact that our results using the standard methodology agree with the predictions of the theory provides a powerful argument for the value of the relevance weighting functions.

References

1. **Angione, P.V.** 1975. "On the Equivalence of Boolean and Weighted Searching Based on the Convertibility of Query Forms." *Journal of the American Society for Information Science.* 1975 March-April; 26: 112-124.
2. **Salton, G.** 1975. *A Theory of Indexing,* Regional Conference Series in Applied Mathematics, No. 18, Society for Industrial and Applied Mathematics, Philadelphia, PA. 1975.
3. **Sparck Jones, K.** 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation.* 1972; 28: 11-21.
4. **Sparck Jones, K.** 1973. "Index Term Weighting." *Information Storage and Retrieval.* 1973; 9: 619-633.
5. **Robertson, S.E.** 1972. Letter, *Journal of Documentation,* 1972; 28: 164-165.

6. **Robertson, S.E.** 1974. "Specificity and Weighted Retrieval." *Journal of Documentation*, 1974; 30: 41-46.

7. **Barkla, J.K.** 1969. "Construction of Weighted Term Profiles by Measuring Frequency and Specificity in Relevant Items." Presented at the Second International Cranfield Conference on Mechanized Information Storage and Retrieval Systems, Cranfield, Bedford: 1969.

8. **Miller, W.L.** 1970. *The Evaluation of Large Information Retrieval Systems with Application to Medlars.* Ph.D. Thesis, University of Newcastle. 1970.

9. **Miller, W.L.** 1971. "Probabilistic Search Strategy for Medlars." *Journal of Documentation.* 1971; 27: 254-266.

10. **Sparck Jones, K.** 1975. "A Performance Yardstick for Test Collections." *Journal of Documentation*, 1975; 31: 266-272.

11. **Barker, F.H.; Veal, D.; Wyatt, B.K.** 1972. "Towards Automatic Profile Construction." *Journal of Documentation.* 1972; 28: 44-55.

12. **Yu, C.T.; Salton, G.** 1976. "Precision Weighting—An Effective Automatic Indexing Method." *Journal of the Association for Computing Machinery*, 1976; 23: 76-88.

13. **Salton, G.; Wong, A.; Yu, C.T.** 1976. "Automatic Indexing Using Term Discrimination and Term Precision Measurement." *Information Processing and Management*, 1976; 12: 43-51.

14. **Salton, G.** (Ed.) 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, N J : Prentice-Hall. 1971.

15. **Robertson, S.E.** 1976. *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems,* Ph.D. Thesis, University of London, 1976.

16. **Nilsson, N.J.** 1965. *Learning Machines*, New York: McGraw-Hill. 1965.

17. **Cooper, W.S.** 1973. "On Selecting a Measure of Retrieval Effectiveness." *Journal of the American Society for Information Science.* 1973 March-April; 24: 87-100.

18. **Cooper, W.S.** 1973. "On Selecting a Measure of Retrieval Effectiveness. Part 2, Implementation of the Philosophy." *Journal of the American Society for Information Science.* 1973 November-December; 24: 413-424.

19. **Tague, J.M.** 1973. "A Bayesian Approach to Interactive Retrieval," *Information Storage and Retrieval,* 1973; 9: 129-142.

20. **Cox, D.R.** 1970. *Analysis of Binary Data,* London: Methuen. 1970.

21. **Robertson, S.E.; Teather, D.** 1974. "A Statistical Analysis of Retrieval Tests: A Bayesian Approach." *Journal of Documentation*, 1974; 30: 273-282.