A probabilistic model of information retrieval : development and comparative experiments

Part 2

K. Sparck Jones[†], S. Walker[‡] and S.E. Robertson[‡*]
[†]Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG
ksj@cl.cam.ac.uk
[‡]Microsoft Research Limited
St. George House, 1 Guildhall Street, Cambridge CB2 3NH
{ser, sw}@microsoft.com
[*]also Department of Information Science, City University, London.

January 2000

Published in *Information Processing and Management*, 36, 2000, 809-840.

Abstract

The paper combines a comprehensive account of the probabilistic model of retrieval with new systematic experiments on TREC Programme material. It presents the model from its foundations through its logical development to cover more aspects of retrieval data and a wider range of system functions. Each step in the argument is matched by comparative retrieval tests, to provide a single coherent account of a major line of research. The experiments demonstrate, for a large test collection, that the probabilistic model is effective and robust, and that it responds appropriately, with major improvements in performance, to key features of retrieval situations.

Part 1 covers the foundations and the model development for document collection and relevance data, along with the test apparatus. Part 2 covers the further development and elaboration of the model, with extensive testing, and briefly considers other environment conditions and tasks, model training, concluding with comparisons with other approaches and an overall assessment.

Data and results tables for **both parts** *are given in Part 1. Key results are summarised in Part 2.*

Keywords: information retrieval; retrieval theory; probabilistic model; term weighting; experiments

In Part 1 we presented the foundations of our probabilistic model and its treatment of basic collection data, with test results showing the performance outcomes for these initial instantiations of the model with our test collections. We now continue the model development and testing. Results for the experiments described in this part are given together with the *previous ones in the Appendix tables to Part 1; however, selected results are repeated here in Part 2, as and when they are discussed in the text. Section 10 contains summary tables of key results from both parts. Any formulae from Part 1 referred to in Part 2 are repeated in Table 9. As in Part 1, the technical report from which this paper is drawn (Sparck Jones, Robertson and Walker 1998) is henceforth referred to as Sparck Jones, Robertson and Walker.*

Table 9: Formulae from Part 1

Equation no.	Formula
3	$W(A_i = a_i) = \log \frac{P(A_i=a_i \vert L)P(A_i=0 \vert \overline{L})}{P(A_i=a_i \vert \overline{L})P(A_i=0 \vert L)}$
4	$\mathit{MS\text{-}BASIC} = \sum_i W(A_i = a_i)$
5	$w_i = \log \frac{p_i(1-\overline{p_i})}{\overline{p_i}(1-p_i)}$
6	$CFW = \log \frac{N}{n_i}$
8	$RW = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)}$

4 Data (continued)

4.5 Term frequencies and weighting

Term incidence data is the most salient for exploitation; and it is important because at least in its most basic form, without relevance annotation, it comes with any file. We can moreover hope, if not expect, to be able to have some relevance incidence data to use as well. But there is other information about term behaviour which may also be available and useful for interpreting the model. In particular, the information used so far assigns different values to terms only according to their distribution across document files, and does not, for a particular term, distinguish one file document containing it from another.

The natural further interpretation of the model is therefore to exploit information about term frequencies within documents, if this can be supplied in initial descriptions. We require some way of modelling term frequency, and (as with the previous data) of relating this variable to relevance. Term frequencies within documents have in the past been modelled using Poisson distributions (Harter 1975); the particular model proposed here is a development of Harter's model (Robertson and Walker 1994). As before, the assumptions from which the model starts are clearly over-simplifications, but may help us develop a useful approach to retrieval.

We assume first that each term is associated with a *topic* (the idea or concept conveyed by that term), and that a document may be *about* the topic or not. That is, for each such topic, there is one set of documents about it and another (its complement in the file) that is not about it. We also allow, however, that the use of the term in text has some unpredictability about it; an author writing about the topic in question may use the term to a greater or lesser extent. Furthermore, an author not writing about this particular topic may refer to it in passing. Since we do not know which documents are about the topic and which not, the distribution of within-document term frequencies that we observe is a mixture of two distributions, one in each of the two sets.

The basic assumption here is that both these distributions are Poisson. We may see this distribution as arising from a very crude language-generation model: if the author is stepping through the possible word-positions in a document, and choosing words to fill them, and furthermore if (a) the probability of choosing the term in question is fixed, and (b) the documents are all of equal length, a Poisson distribution of within-document frequencies will result. Clearly we assume different probabilities for those documents that are about the topic of the term, and another for the others – hence the mixture of two Poissons. The Poisson assumptions also really only make sense if all documents are of equal length. We assume this for now, and return to the document length question later. Furthermore it should be noted that the fit of the two-Poisson mixture to term-frequency data is not in general very good; at best, this is a first approximation.

A Poisson distribution is defined by a single parameter, the mean: but the two distributions are clearly likely to have different means. A complete description of the mixture also requires a third parameter, representing the proportions of the two types of document in the collection. Thus this mixture model has three parameters for each term.

The property of being about the topic or concept referred to by a term is called *eliteness* for the term in (Harter 1975). We denote this E, so E_i means "elite for term t_i", and $\overline{E_i}$ means "not elite for term t_i". TF_i is the frequency of term t_i in the document under consideration. As before, we may drop the suffix. The Poisson distribution assumptions will give us formulae for $P(TF|E)$ and $P(TF|\overline{E})$, in terms of each of the two Poisson means. The same formulae will cover the case $TF = 0$, i.e. the term is absent.

We can also define the probabilities for eliteness given likedness, namely $P(E|L)$ and $P(E|\overline{L})$. The basic assumption (Robertson, van Rijsbergen and Porter 1981) is now that TF depends directly on eliteness only, so that the relationship between TF and likedness is through eliteness. This relationship is expressed by means of two equations, one involving L:

$$P(TF|L) = P(TF|E)P(E|L) + P(TF|\overline{E})P(\overline{E}|L)$$

and a second, similar one involving $P(TF|\overline{L})$. The probabilities of the type $P(E|L)$ imply two further parameters per term, making a total of five.

Now referring back to formulae 3 and 4 in Section 2, Part 1, we see that the event $A_i = a_i$ may be interpreted as $A_i = TF_i$ or just as TF_i, and $A_i = 0$ as "term t_i absent". We may therefore express $W(TF_i)$ in formula 3 as a function of the Poisson distribution parameters and such quantities as $P(E_i|L)$.

The resulting formula is complex. This is not so much a question of algebraic complexity (although that is the case), as complexity of interpretation and estimation. Since eliteness is as invisible as relevance, none of the five parameters can in general be directly estimated. However, in Robertson and Walker (1994), the behaviour of this formula is examined, and a much simpler formula which has similar behaviour is proposed and tested. The simpler formula is as follows

$$W(TF_i) = \frac{TF_i(k_1 + 1)}{k_1 + TF_i}\, w_i \tag{9}$$

Here w_i is the usual presence weight of term t_i (formula 5, Part 1); k_1 is a constant, discussed below. The behaviour of this simple formula, mimicking the complex one, is that (a) it is zero for $TF_i = 0$, (b) it increases monotonically with TF_i, and (c) it has an asymptotic limit.

When $TF_i = 1$, the weight is just the usual presence weight w_i; additional occurrences of t_i increase its contribution to the score, but there is an absolute limit on how much they can add. This asymptotic limit should in fact be the weight that would be associated with eliteness (if we could *know* whether a document was elite for the term or not). This fact provides the justification for multiplying the TF component by a binary presence weight function (which in fact gives us a familiar kind of $TF * IDF$ hybrid). Of course, again, we have no direct knowledge of the binary presence weight for eliteness. However, using the usual term presence weight w_i gives us a plausible approximation.

The constant k_1 determines how much the weight reacts to increasing TF. If $k_1 = 0$, the weight reduces to the term-presence weight only; if k_1 is large, the weight is nearly linear in TF. It may be regarded as a tuning constant, to be adjusted after experimentation with the particular database. In TREC, we have found values in the range 1.2–2 to be effective. This small range implies that the effect of TF is highly non-linear, i.e. after say 3 or 4 occurrences of a term the impact of additional occurrences is minimal. (See TR446 (1998, Section 8) for a discussion of the issues around discovering such values.)

Exploiting Poisson ideas thus means we have a way not only of bringing two separate types of information about terms and documents together, but of capturing the significance of different frequencies for terms in a single document in relation to term behaviour across the file: a document has a higher probability of relevance not simply if a term is frequent in it, but is unusually frequent given the number of documents in which it appears. Further, all of the argument works with any of the earlier instantiations of w_i, for example the collection frequency weight CFW or the relevance weight RW (formulae 6 and 8, Part 1, respectively). These instantiations are discussed in the next section, after considering document length.

4.6 Document lengths and weighting

While it is wholly plausible to take term frequencies into account, the development of the formulae so far has tacitly assumed that all documents are the same length; and indeed the Poisson approach assumes constant length. In practice documents are not merely not all the same length, they may vary widely in length; and it is clear that one document containing a term t should not be preferred to another because t is more frequent in the former than the latter if this is simply because the first document is twice as long as the second. Of course documents may vary in length for different reasons. But if we make one assumption, again of a rather simple but not unreasonable kind given the nature of the retrieval task, we can extend our model interpretation to deal with varying document length.

The simplest assumption is that where there are two documents about the same topic but of different lengths, this is just because the longer is more wordy. When closely examined from a linguistic point of view, as embodying a model of discourse, this is a very crude assumption: it implies wordiness is attributable merely to repetition rather than greater elaboration etc. But as retrieval normally deals with topic description at a a fairly general level, it may be sufficient to equate refinement with prolixity. On this assumption it is appropriate to extend the model interpretation to normalise term frequency by document length.

A simple normalisation (dividing TF by DL) would have the effect of giving the same score to a document of length DL in which a term t occurs TF times, as to a document of length $2DL$ in which the same term occurs $2TF$ times. But the crudity of the assumption is likely to lead to a bias in the above normalisation. That is, the $2DL$ document is unlikely to require a *smaller* score than the DL document, and it may be justifiable to give it a larger one (e.g. if wordiness suggest greater elaboration rather than just repetition). The slightly more complex normalisation suggested below, a mixture of no normalisation at all and the above simple normalisation, allows for this.

Another consideration is how document length may be measured. One could make many suggestions (e.g. word types or tokens, with or without stopwords, or simply characters). It probably does not matter much which is used, but it is appropriate to introduce some uniformity of scaling by relating document length to the length of an average document (in the same units). This will ensure that a document of average length will get the same score after document length normalisation as it had before.

The simple normalisation factor would therefore be $NF = \frac{DL}{AVDL}$. The mixed normalisation factor would be $NF = ((1-b) + b\frac{DL}{AVDL})$, with another tuning constant b, between 0 and 1, discussed further below. Considering the TF component of the TF formula 9 above, after normalisation we have

$$\frac{\frac{TF_i}{NF}(k_1+1)}{k_1 + \frac{TF_i}{NF}} = \frac{TF_i(k_1+1)}{k_1 * NF + TF_i}$$

Hence the weight becomes

$$W(TF_i) = \frac{TF_i(k_1+1)}{k_1 * ((1-b) + b\frac{DL}{AVDL}) + TF_i}\ w_i \tag{10}$$

If the new tuning constant b is set to 1, the simple normalisation factor is used (corresponding to an assumption of pure verbosity). Smaller values reduce the normalisation effect. Experiments with the TREC collection suggest a value of around $b = 0.75$ is good [1].

In order to simplify the presentation of equation 10, we replace $k_1 * NF$ with K, as follows:

$$W(TF_i) = \frac{TF_i(k_1+1)}{K + TF_i}\ w_i \tag{11}$$

where $K = k_1 * ((1-b) + b\frac{DL}{AVDL})$

The assumptions made here about the nature of document length differences are of course not the only possible ones: the most obvious other situation is where length relates to multi-topicality. We return to this in Section 5.

4.7 Instantiations

The last weighting function, 11, encapsulates the way terms gain value within the probabilistic framework from different types of information. It may be instantiated with or without relevance information, as suggested in the previous section. We will give these two instantiations different names. First, when using just the collection frequency weight CFW (equation 6, Part 1), without relevance information, the combined weight is

$$CW = \frac{TF(k_1+1)}{K + TF}\ \log\frac{N}{n} \tag{12}$$

(with a corresponding matching score $MS\text{-}CW$, the sum of the combined weights of the matching terms).

[1] We use the name b rather than the more obvious k_2 for compatibility with other papers, where b is used for this purpose while k_2 is used for something else.

Second, with relevance information, using the relevance weight RW(equation 8, Part 1), the combined iterative weight (named to mark the fact that getting and using relevance information is an essentially iterative process) is

$$CIW = \frac{TF(k_1+1)}{K+TF} \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{13}$$

(again with a corresponding matching score $MS\text{-}CIW$).

In both of these weighting formulae, we have the two tuning constants k_1 and b, as already explained.

4.8 Long queries

The weighting formulae just given would be directly suited to many retrieval applications, namely those where initial queries are quite simple, consisting just of a few words. However, a further refinement of the formulae may be devised to cover the possibility that terms are repeated in the query itself, as would naturally occur if a previously-known document was used as a starting query. The theoretical basis for the refinement is very similar to that for the inclusion of within-document term frequency: a model based on a mixture of Poisson distributions, but applied to the set of queries rather than to the set of documents. As with documents, the justification would be a simple model of a language-generation process applied to the process of writing queries. This is not a strong justification, but may be suggestive of how to proceed.

In principle, this leads to a QTF component similar to the TF component above, but with its own k tuning constant analogous to k_1. Experiments with TREC suggest that the higher the value of such a k, the better (in contrast with the TF component). But as indicated in Section 4.5, large k corresponds to an almost-linear function of term frequency, in this case of QTF. The formulae presented below are therefore based on a simple linear relationship, and no additional tuning constant is needed.

As before, there are two formulae, representing the situation without and with relevance information. The two weighting functions are called $QACW$ (query adjusted combined weight) and $QACIW$ (query adjusted combined iterative weight):

$$\begin{aligned} QACW &= CW * QTF \\ &= \frac{TF(k_1+1)}{K+TF} \; QTF \; \log \frac{N}{n} \end{aligned} \tag{14}$$

and

$$\begin{aligned} QACIW &= CIW * QTF \\ &= \frac{TF(k_1+1)}{K+TF} \; QTF \; \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \end{aligned} \tag{15}$$

Again the weights should be summed over matching terms for the matching scores $MS\text{-}QACW$ and $MS\text{-}QACIW$ respectively.

4.9 Frequency experiments

The introduction of term frequencies leads to a large number of new performance comparisons, when both one-off and iterative searching are taken into account and term frequencies in queries as well as documents are covered. These comparisons only apply directly to the TREC T741000X collection results shown in Table 6 in Part 1, but have indirect connections with the older collections in showing the need to respond to the consequences of using full text as opposed to short document surrogates.

The first comparison is simply one without the use of relevance information, i.e. between plain *CFW* and full *CW* weights, and without query adjustment. Table 10 shows very large performance improvements regardless of whether Long, Medium or Very short queries are involved, and for both Rec30 and Doc30. The gain is more than Dramatic, i.e. $CW + >>>> CFW$, and the significance tests are correspondingly good. Thus these TREC data experiments support the early theoretical arguments for using term frequency information and add to the practical evidence for its value. Both have been notable in the SMART work (Salton 1975, Salton and Buckley 1988), and term frequency weighting has become a general strategy in the TREC Programme. The experiments at the same time confirm the value of frequency weighting for full text, which earlier tests only with abstracts or small sets of longer texts could not convincingly do. It is also noteworthy that simply incorporating term frequencies (suitably normalised), for these full text cases, is not merely better than the most favourable predictive *RW* relevance weighting, using all documents, but than retrospective relevance weighting. Thus the difference, for either performance measure, is at least Striking, i.e. $CW >>> RW\ retro$, and the difference is significant.

Table 10: Extract from Table 6

	Doc30			Rec30			AveP		
	L	M	V	L	M	V	L	M	V
CFW	.07	.15	.17	.04	.10	.17	.03	.07	.12
CW	.41	.40	.40	.32	.32	.34	.23	.23	.24
QACW	.50	.44	.40	.44	.37	.34	.32	.27	.24
QACIW retro	.54	.48	.44	.49	.43	.38	.35	.30	.27
RW retro	.30	.26	.19	.25	.24	.20	.18	.17	.13
QACIW pred all	.54	.48	.43	.48	.42	.37	.35	.30	.27
RW pred all	.29	.26	.19	.25	.24	.20	.17	.17	.13
QACIW pred top 3	.51	.46	.42	.46	.39	.35	.33	.28	.25
RW pred top 3	.16	.21	.18	.13	.18	.18	.08	.12	.12
QACIW pred rel in 10	.52	.46	.42	.46	.40	.36	.33	.28	.25
RW pred rel in 10	.21	.23	.18	.17	.20	.18	.12	.14	.12

When query adjustment, *QA*, is applied in *QACW* versus *CW* there is, not surprisingly, further gain for the Long requests, with $QACW >>>> CW$ on either measure, and also for Medium where $QACW >> CW$, but nothing for Very short requests. The informal comparison is borne out by the significance test results. But the lack of improvement for the V requests is hardly surprising with only 4 terms per request; and since the general evidence is in favour of rather than against using frequency information for queries when it is available we have adopted query adjustment as the default in tests with other strategies.

Thus if we now consider the combination of term frequency and relevance information, under the label *QACIW* in Table 10, and compare the corresponding runs for *QACIW* and for *RW*, where relevance information alone is used, we find very marked gains from using frequencies. Thus for the retrospective case the performance improvement on either measure is more than Dramatic, i.e. $QACIW\ retro\ + >>>>\ RW\ retro$, and the same holds for the predictive cases using all or top 3 relevant documents: i.e. $QACIW\ pred\ all\ + >>>>\ RW\ pred\ all$, and $QACIW\ pred\ top\ 3\ + >>>>\ RW\ pred\ top\ 3$. These large differences are also independent of request form, and are highly significant. The same holds for the comparison using *pred rel in* 10. Thus these runs very clearly show the advantages of a more refined treatment of the primary data about term occurrences within documents, even when relevance information is also available.

At the same time, checking runs for the TREC collection in order to compare the two treatments of the yardstick formula, without and with 0.5, for $QACIW$ give rather different results from the analogous ones for the older collections using RW (there is little point in trying both RW yardsticks for TREC when performance for RW is so low compared with that for CIW). Thus for T741000X, $QACIW\ retro * = QACIW\ retro$. But this is not surprising, given the much larger numbers of relevant documents available for prediction.

It is more important to compare $QACIW$ with $QACW$ for the effect of relevance information within the combined weight context and also, given the potential advantages of iterative weights, to consider more carefully the effects of variations in the amount and/or quality of relevance information. The runs shown in Table 10 confirm the value of extensive relevance information, with both Rec30 and Doc30 measures, though the gains are perhaps not as large as might be expected: thus in the weakest case, the improvement is no more than Noticeable, i.e. $QACIW\ pred\ all > QACW$. Further, when only the top 3 relevant documents are used for prediction, there is no gain at all in most cases: thus the *general* conclusion for the top 3 runs cannot be other than $QACIW\ pred\ top\ 3 = QACW$. The difference in comparative behaviour for $QACIW$ *pred all* and *pred top* 3 is supported by the significance test results.

Indeed even with retrospective relevance weighting, the gains from relevance information are not enormous. Thus while the difference between having the best relevance information and not having any at all is material for the Long and Medium request forms, this is not an impressive difference, and the difference is smaller for the Very short requests, so the overall picture is only $QACIW\ retro > QACW$. Further, as all the foregoing implies, there is also little difference between $QACIW$ retrospective and $QACIW$ predictive, whether using all relevant or only top 3 for the latter. Thus $QACIW\ retro = QACIW\ pred\ all$, though $QACIW\ retro > QACIW\ pred\ top\ 3$. Again, this informal view is endorsed by the significance tests.

Altogether these results suggest that while there are large performance gains to be made from using term frequencies, those to be further made from adding in relevance information may not be so large, certainly on the rather strong view we are taking of worthwhile performance differences. It is possible that the rich data supplied by full texts means that the extra power we would expect to gain from exploiting relevance data in addition to ordinary frequency data is largely preempted, explaining both why the performance ceiling represented by the retrospective case is not relatively higher and prediction gets much closer to it than with the old collections. But as there are nevertheless some performance gains, it is worth exploring the effect of alternative prediction bases, as in Section 4.2 in Part 1. We report in Table 6 in Part 1 the results of 'rel in 10' experiments (i.e. using whatever relevant documents are found in the best matching 10 documents); we have also compared these relatively favourable cases with using 3 relevant documents drawn at random from the full set ('rand 3'), replicating the purely quantitative base studied in earlier experiments, but in a more stringent way through drawing a smaller sample from the larger TREC relevance set. We consider later, in Section 5.3, the case where best matching documents on a first pass are deemed relevant, whether they actually are or are not so (the runs labelled 'blind').

The results given in Table 6 show that, regardless of request form and on both performance measures, while the predictive all case is Noticeably better than the other prediction bases except for the Very short requests, there is no difference between the others. More specifically, if we exclude the slightly unexpected results for 'rand 3', which performs better than might be expected for the Very short requests, we find that $QACIW\ pred\ all > QACIW\ pred\ rel\ in\ 10$, as we earlier found $QACIW\ pred\ all > QACIW\ pred\ top\ 3$; the similarity in performance for the two small bases presumably reflects the fact that there are sometimes more than 3 relevant documents in the top 10, sometimes less. The difference is also, as before, significant. But unfortunately when we compare having limited relevance information with having none at all, there is either little or no performance gain. That is, as noticed above, while having an unrealistically large supply of relevance information, as in 'all', does improve performance compared with $QACW$, so $QACIW\ pred\ all > QACW$, having only a little relevance information is not guaranteed to contribute anything useful: $QACIW\ pred\ top\ 3 = QACW$, and the same for rel in 10. The significance data in Table 8 in Part 1 agree for the top 3 case, though the difference for rel in 10 is significant. These results for $QACIW$ compared with $QACW$ hold regardless of request form, which appears surprising. But presumably, with only a few relevant documents, the frequency behaviour of terms in the document set as a whole is the dominant factor and the relevance information is not very discriminating. The explanation for the better 'rand 3' results than might have been expected may be that the random relevant documents give better coverage of the relevance sample space than the top ranked ones.

5 Elaboration

The core model we have presented can be developed in various ways to take account of further sources of information. Exploiting these allows a more refined treatment of query-based indexing and matching and hence, potentially, better discrimination between wanted and unwanted documents. Some of the notions involved have indeed, in our own and other analogous tests, been shown to improve performance; the first we discuss, query expansion, is well established as very effective. However while the elaborations we consider in this section are all well motivated, their very complexity may make it more difficult to choose an appropriate instantiation of the general idea in question, and typically imply a requirement for some application-specific training. We consider questions of training in Section 8 later, and more fully in TR446 (1998). But the practical difficulties of obtaining sufficiently varied training data in the form of a range of test collections mean that it is impossible yet to draw very firm conclusions about some of the model elaborations we consider.

5.1 Query expansion from relevance information

It is evident that restricting query development just to providing term weights is not taking advantage of all the means of improving performance that are on offer. Specifically, restricting relevance feedback to reweighting is a very limited way of exploiting the information that retrieved and judged documents supply. Thus as Rocchio (1965) and Ide (1968) early saw, it is natural to consider terms in retrieved relevant documents as possible additions to a query for further searching (as also, perhaps, to eliminate existing query terms found not to be in relevant documents). While under the most abstract view of retrieval a query may be seen as consisting of all the terms in the file vocabulary, with positive or negative weights, we here consider only those terms which seem to have some connection with the query. Thus a case can be made for seeing terms not already in the query (as earlier), but nevertheless in documents marked as relevant to it, as potential query members.

It does not, however, follow that the candidate expansion terms, i.e. all those occurring in at least one relevant document, should simply be added to the query, even if they are then weighted using relevance information before searching. It may be better to select from the set, depending on the environment conditions that apply in the given application, for instance to avoid 'blowing up' an initial short query with an enormous number of terms that are candidates because documents are long. The constraints different conditions impose on selection are in fact not well understood (see Section 5.2 below). But under the assumption that selection is required, an obvious procedure for selection is to rank candidate terms and apply a cutoff to this ranking.

It might appear that the appropriate basis for ranking is the same as that for the ordinary term weighting using relevance information. But the question that is answered by a relevance weight ("How much evidence does the presence of this term provide for the relevance of this document?") is not actually the same as the question being asked here: "How much will adding this term to the request benefit the overall performance of the search formulation?". In particular, a very rare term, even though it is a strong indicator of relevance when it occurs, is not likely to have much overall effect.

A specific model for the overall effect is discussed in Robertson (1990). For each candidate expansion term, this considers the distributions of the scores for relevant and non-relevant documents, with the term in question present or absent. The model leads to a formula for a 'selection value' for term t_i, indicating the strength of its overall effect, which may be specified as follows (using the notation introduced in Section 2.6, Part 1):

$$\text{selection value} = (p_i - \overline{p_i})w_i$$

In this formula, the ps are as defined as in Section 2.6, but the ws represent whatever weight will be applied to the term in question. The formula assumes, however, that the term weight is independent of the document – e.g. there is no TF or DL effect. We will therefore interpret this weight as the relevance weight RW or the collection frequency weight CFW, not as any of the combined weights.

In practice, $\overline{p_i}$ is generally very much smaller than p_i, at least for terms that have significant weights w_i, and can in general safely be ignored.[2] Moreover, if p_i is estimated by $\frac{r_i}{R}$, the denominator will generally be the same for all terms under inspection, and can therefore also be ignored for the purpose of ranking terms. These considerations lead to the definition of the following simplified selection value, the offer weight

$$OW = rRW$$

(ignoring subscripts).
Once selected, expansion terms are weighted in the usual way with CIW or $QACIW$ (in which RW of course figures).

[2]To put it another way, any term for which $\overline{p_i}$ is not much smaller than p_i will be rejected because it has small w_i.

The treatment of original query terms requires some discussion. One possibility is to keep these terms in the query anyway, and merely consider the addition of new (different) terms; another is to regard the query terms as candidates for the new query, to be given offer weights and ranked with the rest. Under this latter method, it may occasionally happen that an original query term is dropped. There are arguments for either the former, 'addition', or the latter, 'replacement', approach; for simplicity, we have used the second, replacement, method. Thus although we refer to what we do as "query expansion" (and it will generally enlarge the query), we may in fact lose some of the original terms.

Given that the original query terms are included in the analysis and given offer weights, there is a question about how to deal with the query adjustment discussed in Section 4.8. This component of the formula is not document-dependent, and therefore can be included in the offer weight. The formula used in the experiments reported here is therefore

$$OW = r * QTF * RW \tag{16}$$

where QTF is taken to be 1 for expansion terms.

A compromise between the two methods of treating original query terms would be to give them some prior advantage, but to allow sufficient negative evidence from relevant documents to cause their exclusion. One such method is tried below (Section 5.2).

Earlier experiments with the older collections reported in Sparck Jones and Webster (1980) explored various forms of expansion, but the tests were limited to the C1400I collection, and the conclusion that there appeared to be mileage in combining query expansion and relevance weighting was not followed up. The results for the T741000X collection in Table 11 and Table 6 cover a larger and more systematic range of experiments (though still for only one document file), and combine expansion with the more sophisticated type of frequency-based weighting described in the preceding section. Thus all of these experiments combine expansion with $QACIW$ weighting. We consider below the effects of different amounts of expansion. The tests here refer to moderate expansion, adapted (as seems reasonable) to the initial query request size. Thus we allowed (up to) 32 additional terms for the Long form requests, 24 for Medium, and 16 for Very short. These values, labelled 'exp 32' etc in Table 6, are suggested by a range of experiments as being, if not optimal, at least reliably good for these requests; but the necessity to validate them in this fashion is a very good example of the need for training. For convenience we will refer to these in future as the *default* expansion set sizes, and also assume these sizes in Table 11 and when characterising comparative performance results, unless otherwise stated.

The first set of comparisons is therefore between expanded runs, $QACIW + E$, and unexpanded $QACIW$ runs, for the corresponding prediction bases, namely top 3 and rel in 10. Apart from the Rec30 figures for the Very short requests, where there is no gain from expansion, both Rec30 and Doc30 show at least a Noticeable improvement and typically more, i.e. in most cases $QACIW + E\ pred\ top\ 3 > QACIW\ pred\ top\ 3$ and $QACIW + E\ pred\ rel\ in\ 10 > QACIW\ pred\ rel\ in\ 10$. This also holds for the significance tests, except for the one case of Rec30 for Long requests with top 3, where there is no difference even on the sign test.

Unfortunately, it is not possible with expansion to establish yardstick performance in the same style as for unexpanded requests. Properly, yardstick performance has to *include* the actual choice of expansion terms. But this is a serious combinatorial problem, essentially retro-engineering the perfect request at the level of term choice. We have therefore adopted

Table 11: Extract from Table 6

	Doc30			Rec30			AveP		
	L	M	V	L	M	V	L	M	V
QACIW pred top 3	.51	.46	.42	.46	.39	.35	.33	.28	.25
QACIW + E pred top 3	.54	.51	.43	.48	.45	.35	.35	.32	.25
QACIW pred rel in 10	.52	.46	.42	.46	.40	.36	.33	.28	.25
QACIW + E pred rel in 10	.57	.52	.43	.51	.47	.36	.37	.34	.26
QACIW retro top 3	.52	.46	.42	.46	.40	.36	.34	.28	.25
QACIW + E retro top 3	.58	.55	.48	.52	.48	.40	.39	.36	.29
QACIW retro rel in 10	.52	.47	.42	.46	.40	.36	.34	.29	.25
QACIW + E retro rel in 10	.58	.54	.49	.52	.49	.43	.39	.36	.31

a more limited approach, arguing by analogy with the earlier use of retrospective weighting on the assumption that the given request terms are the proper choice. Thus we assume that the expanded query gives the proper choice of terms, so retrospective performance is about the proper weighting for these. Then for comparisons there are distinct sets of retrospective figures, one for each choice of expansion base, where each comparison is designed primarily, as before, to check the value of reweighting for each term set context. This is a somewhat more heuristic approach than in the previous case, but can still be practically useful.

Following this line of argument we can, at least semi-legitimately, compare retrospective performance with and without expansion. For the default degrees of expansion, and whether via top 3 or rel in 10, and for both Rec30 and Doc30, the expansion yardstick is at least Noticeably better, i.e., $QACIW\ +\ E\ retro\ top\ 3\ >\ QACIW\ retro\ top\ 3$ and $QACIW\ +\ E\ retro\ rel\ in\ 10\ >\ QACIW\ retro\ rel\ in\ 10$. These differences are also significant.

But the more relevant comparisons, for our Half collection, are between retrospective and predictive expanded query performance using, respectively, top 3 and rel in 10, and the default degrees of expansion. For both measures, retrospective performance is Materially better for the former, and at least Noticeably better for the latter except for the Long requests, i.e. $QACIW\ +\ E\ retro\ top\ 3\ >>\ QACIW\ +\ E\ pred\ top\ 3$ and (except for Long) $QACIW\ +\ E\ retro\ rel\ in\ 10\ >\ QACIW\ +\ E\ pred\ rel\ in\ 10$. Table 8 shows that the top 3 comparison differences are also statistically significant. However the differences for rel in 10 are not significant for Medium as well as Long requests.

As just described, query expansion is primarily an automatic strategy, requiring no more user involvement than judging documents for relevance. But of course expansion terms may be presented to the user, in *OW* order, for more participative query reformulation in searching. We consider this in the later Tasks section, Section 7, and in more detail in TR446 (1998).

5.2 Selective or massive expansion?

The question of whether expansion should proceed by selecting a few good terms, or by including every term that occurs in at least one relevant document, or even by including all the terms in the dictionary, is to some extent open. Potentially, every term provides some evidence (positive or negative) concerning the possible relevance of every document, and an appropriate weighting scheme would give each term an appropriate weight. If a term is strictly statistically neutral about relevance, the weighting scheme should give it zero weight, which is actually equivalent to excluding it, but would require no explicit exclusion step.

However, given that the expansion is based on a limited and fixed "sample" of relevant documents, estimating more parameters (i.e. weights) from the sample is likely to result in less accurate estimating of the total score. This argument has been described as the "curse of dimensionality" (van Rijsbergen 1979), and can be formalised mathematically (Robertson and Bovey 1982). Moreover, while the curse of dimensionality can be overcome by a suitable Bayesian prior distribution for the parameters, which would have the effect of damping the response of the estimate to small samples, it is not easy to define appropriate priors.

Empirically, some procedures for query expansion seem to be affected by the curse of dimensionality, in the sense that expanding beyond a certain point degrades performance. Other approaches appear not to be affected, at least to the extent that expansion with hundreds of terms has proved helpful. However the main results here, especially for so-called massive expansion, have been in TREC Programme routing experiments (TREC 1992–1999), where there have been rich requests and large relevance sets for training. Massive expansion was, moreover, found helpful only in earlier TREC evaluations; in later tests with shorter requests and different weighting methods it has not proved superior to more modest expansion (Buckley, Singhal and Mitra 1996). In general, for the methods we apply, the evidence of past tests suggests that limitation is required, and this motivated the selective strategy described above.

At the same time, expansion does not seem to be affected by small changes in the size of the expansion set. The results in Table 6, for T741000X, illustrate predictive expansion for various set sizes for the different request forms, with rel in 10 as the expansion base. These show, for both Rec30 and Doc30, that increasing the expansion set for V requests from 16 to 24 has no effect, and that using 16,24 or 32 for M requests has no effect; with the L requests each step considered individually does not improve performance: only when the two extremes considered, expanding by 16 terms or 48, are compared, is there a Noticeable difference; however continuing to expand as far as 72 has no further value for the Long requests.

One suggestion, in the expansion context, is that the user's original query terms should always be given extra weight, and we have explored this to a limited extent using the following argument. We suppose that the use of a term in the original query is equivalent (as evidence) to its presence in a certain number of relevant documents. These may be supposed to be documents not in the collection, but previously known to the user. Instantiating this argument requires specific assumptions about numbers: we have assumed that the user knows about 20 relevant documents which are not in the collection, and each query term occurs in 19 of them (a fairly strong bias towards query terms). These additional relevant documents figure in the calculations of both the term weight and the offer weight. Table 6 compares predictive performance with and without query term emphasis, for the default expansion sets and using both top 3 and rel in 10 as expansion bases. This very small experiment indicates, however, that there is no gain from emphasising query terms.

5.3 Expansion without relevance information

Expansion using known relevance information has suggested a related search strategy using *assumed* relevance information, which is independent of user intervention (Evans and Lefferts 1995). Thus top-ranking documents from an initial pass are assumed relevant and terms from them are added to the query for the 'real' search. This is for obvious reasons likely to be much less effective than using real relevance information, and may be positively damaging for individual queries with poor initial searches. The utility of this strategy clearly depends on the characteristics of user needs, requests, and document files, but there is some evidence to suggest that the procedure can be beneficial when initial query quality is good. It has certainly become popular in TREC (Sparck Jones 1999c).

Since this *blind* strategy would be a practically convenient one, we explored it for the T741000X collection. The results labelled 'blind' in Table 12 illustrate performance when best matching documents are used for the default degree of expansion, using the 10 best for the Long and Medium form requests, but a more cautious 7 best for the Very short requests. Comparing expansion using either top 3 or rel in 10 with this blind expansion shows some difference in results for the two measures: for Rec30 performance using known relevant documents is sometimes the same as for blind expansion, and at best only Noticeably better; with Doc30 known relevant documents appear more effective. More particularly, expansion using top 3 is not necessarily better than blind expansion, but expansion by rel in 10 is at least Noticeably better than blind expansion, i.e. $QACIW\ +\ E\ pred\ rel\ in\ 10\ >\ QACIW\ +\ E\ pred\ blind$. However this difference is only statistically significant for L and M requests, not V ones.

Table 12: Extract from Table 6

	Doc30			Rec30			AveP		
	L	M	V	L	M	V	L	M	V
QACIW + E pred rel in 10	.57	.52	.43	.51	.47	.36	.37	.34	.26
QACIW + E pred blind	.53	.49	.40	.47	.45	.34	.35	.32	.24
QACIW pred blind	.50	.45	.40	.44	.39	.34	.32	.28	.24
QACW	.50	.44	.40	.44	.37	.34	.32	.27	.24

The second comparison is between no expansion and blind expansion. This also implies no known relevance information for weighting, though of course it is possible to reweight query terms using the data from blind feedback. The comparisons are therefore with $QACIW\ +\ E\ pred\ blind$ against $QACW$, and against $QACIW\ pred\ blind$. These show a gain with blind expansion compared with $QACW$ which is at least Noticeable for Long and Medium requests (i.e. $QACIW\ +\ E\ >\ QACW$), though there is no gain for Very short requests. The same applies to the comparison with and without expansion, i.e. $QACIW\ +\ E\ pred\ blind\ >\ QACIW\ pred\ blind$, except for V. Not surprisingly, as the foregoing implies, reweighting alone using blind feedback gives no improvement, i.e. $QACIW\ pred\ blind\ =\ QACW$. The statistical significance tests confirm this pattern.

Query expansion in general is a broad notion, and one that covers query development before as well as after any searching. It is assumed, probably rightly, that initial queries are typically rather sparse, and could therefore benefit from expansion before searching at all; and many methods of forming fuller queries have been tried, for example taking words or phrases from thesauri. Any expansion method from outside the probabilistic model could, in principle at least, be combined with the probabilistic model, in the sense that new terms could be weighted in the same way that original query terms are. One might however argue that such terms should be given lower weighting status than the original query terms. But this has to be done pragmatically, and hence in some fairly arbitrary way, because at present there is no formal justification in the probabilistic model for any specific way of doing it. There is the further point that the queries most in need of expansion (say 1–3 initial words) are also those for which there is least initial leverage for expansion. This in itself is is a strong argument for using relevance information, since this must provide more leverage.

It is also possible to apply feedback, using search output, in a more informal and less constrained way than in the model-based expansion we have described. Thus output might suggest thesaurus entry points to the user, or otherwise prompt the user to supply further terms. But these strategies could be handled within the model either by treating them as new initial query terms, to be given non-relevance weights or, if output has been assessed for relevance, by giving them relevance weights. Similar problems to the above, concerning the relative value of such new terms compared to the originals, exist in this case. Another open question would be: given relevance information, is it better to expand automatically or with user intervention? This question is further discussed in TR446 (1998).

5.4 Term cooccurrences

The assumption that terms occur independently, on which all of our model development so far has been based, is patently incorrect: given that topics are complex, terms expressing them can be expected to cooccur. Thus in principle, the formulation of probabilities about documents being liked given single terms should be elaborated to cover probabilities about documents being liked given combinations of terms. However allowing for all possible combinations is computationally intractable, and attempting to deal with term dependencies in a more practically acceptable way as in e.g. van Rijsbergen (1977) or Robertson and Bovey (1982) has not delivered notably useful results. The assumption must be that information about term dependencies with respect to whole documents is too weak to be helpful. In any case, because the independence assumptions of our model are conditional on relevance, they actually imply some dependence between terms: if two terms are good for a query and are independent given relevance, they will also be co-dependent in the whole collection. Thus the model may to some extent capture, indirectly, the important dependencies. Further, Cooper's generalisation of the model (Cooper 1995) suggests that our independence model equations are resilient to distortion from at least some types of dependencies.

5.5 Term phrases

In principle the model should cover *compound terms* or phrases (Sparck Jones 1999a), and there is indeed no difficulty about incorporating undecomposable phrases. However dealing with decomposition, i.e. treating both phrases and their members separately within the model presents unsolved difficulties, even though defining phrases by statistical association, as in Buckley, Allan and Salton (1995), might be more compatible with the model than an explicitly syntactic approach. But since all the general evidence, as illustrated by Mitra et al. (1997) and the analysis of TREC in Sparck Jones (1999c), shows that performance gains from phrases, however defined, are not large, the stimulus to tackling the issue of weighting for phrases within the model has been lacking.

5.6 Document levels: passages

Document, as much as terms, may be treated in a more sophisticated way than hitherto assumed, by retrieving component passages rather than whole documents, or by using passage-level information instead of or as well as document-level information for scoring matches. Passage retrieval in its own right has not been tested properly anywhere, through lack of the required relevance information. Combining *global* with *local* information presents difficulties for the probabilistic model, but using passage-level information only to determine a document's matching score does not. Our tests on this have defined passages *dynamically* rather than *statically*, i.e. have used the best scoring passage per document from candidates of different lengths: this is in principle better than using fixed-length passages. However the results given in Table 6, included for completeness, show varied and nowhere interesting results. Thus comparing first retrieval without expansion, using $QACIW$ and weighting via rel in 10, for both Rec30 and Doc30 measures there is no Noticeable difference. The same applies with query expansion, $QACIW + E$, using the default expansion sets. Recent TREC work (Sparck Jones 1999c)cf suggests that using passages round matching query terms to bound the sources for query expansion terms seems to be useful. We have not tested this significantly.

These last three forms of elaboration, namely via dependencies, phrases, and passages, are more fully discussed in TR446 (1998): we have treated them only briefly here since they appear to have little impact on performance. Thus the overall conclusion on elaborations of the core model is that those which do not use relevance information are of little value. But where relevance information is exploited to modify queries, this is a large important data change with a large effect on performance.

6 Environment conditions

6.1 Document properties

The test collections shown in Table 2 in Part 1 are very varied, and the TREC data is internally heterogeneous. The TREC full text documents stimulated the adoption of K, but the variation in length is not (it appears) enough to invalidate the verbosity hypothesis. In general, the TREC Programme tests, as illustrated by the City papers in TREC (1992–1999) and by the surveys in Sparck Jones (1999b, 1999c), suggest that the probabilistic approach is quite robust under considerable collection heterogeneity, and this is confirmed by the recent tests with the TREC Very Large Collection track (Hawking and Thistlethwaite 1998), which contains a mixture of document types and includes USENET news. But there has been no systematic testing for the effects of really different types of material. It should also be noted that the model has been successfully applied to spoken documents, initially on a small scale (Sparck Jones et al. 1996, Jones et al. 1996) but recently on a larger one (Johnson et al. 1999), though transcription is far from perfect and broadcast news is a genre all its own.

6.2 Request properties

Even for the routine adhoc retrieval task with which we have been concerned so far, there may be great variation in requests as expressions of user needs, and hence in the term composition of initial queries. This variation includes not only e.g. differences in concept generality and other content properties that are not directly acessible though they may be inferred in the model from term incidence frequency; it also covers differences in care and elaboration which are also not directly accessible but may be inferred from the number of terms provided, and are indeed exploited for retrieval in the basic postcoordinate approach.

In general, very short initial queries, while working better than longer ones with simple weighting methods, provide little leverage for developing better final queries. With more sophisticated weighting schemes performance for longer requests is better. Encouraging end-users to provide good starting requests (and hence queries), for example for searching the World Wide Web, is further examined under Open Issues later. However we note here that it is one of the major strengths of the probabilistic model that the methods of term weighting drawn from it lead to relative performance improvements even where absolute performance, through poor queries, is not impressive. This is illustrated by an overall comparison between performance for Long, Medium and Very short requests as shown in Table 6. Thus the same strategies as improve performance for L and M improve it for V, except that the first two gain much more from expansion (as well as from query adjustment). The TREC–5 and –6 tests also supply some confirming data here. The former offers 'Short' (in fact Description field only) versus 'Long' (full topic) comparisons for rather difficult requests, the latter 'Very short' (Title field only), 'Short' and 'Long'. Unfortunately Title terms were not necessarily included in the Description field, and for TREC–6 in particular the Title had its own considered character; the comparisons illustrated in Sparck Jones (1999b, 1999c) thus show 'Very short' queries performing better than 'Short'. But these comparisons also, more importantly, show City performance with automatic searching as among the best for any of the request forms, though the absolute performance levels vary.

6.3 Languages

The results we give are for English. There is no reason in principle why the model should not carry over to other languages, and indeed the test results obtained by City for the TREC Chinese material discussed in Smeaton and Wilkinson (1997) suggest that it can, even for a language with a rather different type of basic term.

7 Tasks

7.1 Interactive searching

We have already considered the natural extension of adhoc searching to iteration, especially in the interactive session mode. Thus we have envisaged offering the user candidate expansion terms for review and selection. It is clear, also, that it is important as a practical matter to encourage the user to make the relevance judgements on which performance improvement depends. But there are general questions about the relative authority of user and system information, and about whether interactive searching is a genuinely distinct task, i.e. about whether the probabilistic model can be fully integrated with the way the user sees and does retrieval. The Okapi system using the model is an operational one, and has been used in a series of tests with real users with real information needs (Okapi 1997). But the results from detailed studies of interactive searching within TREC have been quite limited.

7.2 Text extraction and summarisation

Clearly, the information about word value on which the probabilistic model depends could be applied to help identify key material in a longer text, whether at the passage or phrase level. This key material could be used to form summaries, for a variety of purposes, as illustrated by the selective treatment of passages in Salton et al. (1997) on the one hand, and the use of phrase sets as 'mini-summaries' for a variety of full text preview or other assessment and browsing functions as considered by Hand (1997) and Mani and Maybury (1997). This is an area of active research in general, and one to stimulate model development and application, for instance in its foundational definitions with respect to relevance and its practical utility.

7.3 Routing and filtering

The 'pure' specification of *filtering* differs distinctively from the adhoc case. In the latter all documents are considered in relation to a query, in the former a single document in relation to each query. There is also more opportunity with filtering for long-term learning. Filtering is thus a natural field for application of the model. Until recently, however, it has been studied in TREC only in a limited form, with 'routing' defined as retrieval and ranking for a new batch of documents. Not surprisingly, the probabilistic model has performed well here (TREC 1992–1999)see City papers in, exploiting iterative query optimisation on a document training set. Here results shown in Table 13 can be used to illustrate the value of rich training data. Thus if we compare $QACIW\ +\ E\ pred\ rel\ in$ 10 (something like the best reasonably achievable with relevance information) with $QACW$ (the best without), the former is Strikingly better than the latter on Long and Medium requests, though the difference is only Noticeably so on V.

Table 13: Extract from Table 6

	Doc30			Rec30			AveP		
	L	M	V	L	M	V	L	M	V
QACIW + E pred rel in 10	.57	.52	.43	.51	.47	.36	.37	.34	.26
QACW	.50	.44	.40	.44	.37	.34	.32	.27	.24

Learning under the model has also, more recently, been applied to yes/no retrieval in a more properly defined TREC filtering task (Walker et al. 1998). However tackling the true filtering task also requires development of the model, since it is now necessary to use a scoring method that allows the specification of an absolute retrieval threshold. One way to do this is to make the score reflect the absolute probability of relevance: this point is reinforced by the use of utility measures for filtering evaluation in TREC (Lewis 1997), which may be interpreted in terms of a probability-of-relevance threshold. The probabilistic model has now been developed in an appropriate way (Robertson and Walker 2000), but as the topic is a large one and testing has not been on the same data as used in this paper, we will not go into further detail here.

7.4 Categorisation

From some points of view filtering is categorisation, but there are other traditional forms of categorisation, like assigning subject headings to documents. Though the model has been presented using simple natural language terms, it can also be applied to documents and requests indexed with such descriptors, as long as they can be treated within a coordination framework regardless of any structural relations between them. Probabilistic approaches may also be candidates for assigning category labels, as in Biebricher et al. (1988), or for forming categories. However as there is no direct appeal to relevance in such situations, the model would have to be reformulated using other primitives, and the precise way in which these indexing-oriented notions could be related to the grounding notion of relevance remains unclear, though there are possible starting points in Maron and Kuhns (1960) and Robertson, Maron and Cooper (1982).

The potential task applications for the probabilistic model are considered in more detail in TR446 (1998).

8 Training

Automated training is a significant feature of many approaches to retrieval, ranging from modest system *tuning* at one end to full-blown *learning* at the other. Training may apply at a system-wide level, or more specifically, as in using relevance feedback to modify an individual query. As the probabilistic model is applied in more situations there will more requirement for effective training. In some cases, notably relevance feedback, the model lays down what to do. In others, but less directly, it offers opportunities, discussed in more detail in TR446 (1998), that call for further study.

Thus with respect to information need, relevance judgements provide evidence about particular needs: but it is not clear whether it is possible to generalise about needs (and about the queries that are their representatives). The probabilistic model may also be applied, as in Maron and Kuhns (1960) and Robertson, Maron and Cooper (1982), to learning about documents, and specifically to learning about the needs to which an individual document may be relevant: the problem with doing this effectively in practice is whether the necessary large training data is available. Further, while training for both queries and documents requires abstraction through terms, i.e. via their specific term descriptions, there is no lead from the model to training about terms independent of their contexts – indeed their context of use is all-important. Some general properties of terms, notably collection frequency, nevertheless figure in the model, and it might therefore be possible to develop the model further towards learning about the general value of terms.

With respect to specific learning techniques, it is possible to regression analysis within the framework of the probabilistic model, and this has already been done in a conventional way for both indexing and searching by Fuhr and Buckley (1991). However since the variable to be learned, relevance, is binary, logistic regression is more appropriate and fits better with the model (Robertson and Bovey 1982, Cooper, Chen and Gey 1994): see Robertson and Walker (2000). Naturally, with the large quantities of data now available, more informal ways of deriving useful relationships can also be pursued as in the work done by Singhal, Buckley and Mitra (1996) on refining document length normalisation methods. However there are potential problems, repeatedly noted in TREC, about learning on past data where complex models with many parameters may lead to overfitting to quirks of the training data, and hence loss of predictive power.

9 Comparisons

The probabilistic view of information retrieval has inspired a number of very different approaches, models, methods and techniques. It is also true that many of the specific methods discussed in this paper have been used in the context of other, non-probabilistic (or not explicitly probabilistic) approaches. Many comparisons could be made, at the level of theories, models, techniques, experimental results, or whatever, between the ideas discussed here and those reported by other researchers. In this section, we make a small selection of such comparisons, concentrating on some major alternative or complementary views, and on ideas which may shed light on the foregoing discussions. (For further background see Sparck Jones and Willett (1997), and also the recent review of TREC in Sparck Jones (1999c).)

9.1 The vector space model

By far the best-developed non-probabilistic view of IR is the vector space model (VSM), most famously embodied in the SMART system (Salton 1975, Salton and McGill 1983). In some respects the basic logic of the VSM is common to many other approaches, including our own: see the discussion of properties (attributes) in Section 2, Part 1. It is also true to say that the VSM is hospitable to other theories, and indeed there are implementations of probabilistic ideas within the VSM. However, the point of departure for the VSM is that the attributes are to be regarded as the axes of a space, and that the required measure of association (e.g. between documents and queries, what we have in this paper described as the matching score) should be a distance measure in this space. In principle at least, this motivation is very different from the 'probability of relevance' motivation which informs the present paper.

In practice the difference has become somewhat blurred. Each approach has borrowed ideas from the other, and to some extent the original motivations have become disguised by the process. Two examples may be given. The idea of relevance feedback originated in the context of the VSM (Rocchio 1965), but also fits very well into the probabilistic approach, as has been seen. Second, the experimental success of the form of document length normalization described in Section 4.6 inspired the SMART system researchers to rethink their own document length normalization (Singhal, Buckley and Mitra 1996).

This mutual learning is reflected in the results of successive rounds of TREC. Typically SMART, Okapi and some of the other systems discussed below are among the best-performing systems with relatively little to choose between them (at least compared to the range of performances represented). It may be argued that the performance differences that do appear have more to do with choices of the device set used, and detailed matters of implementation, than with foundational differences of approach.

9.2 Probabilistic indexing and a unified model

The first explicitly probabilistic model in IR was due to Maron and Kuhns (Maron and Kuhns 1960, Robertson, Maron and Cooper 1982). While it is concerned with probability of relevance, it starts from the opposite end from us: user queries are assumed fixed, but document indexing requires optimization. No real experiments have ever been done with this model.

An attempt has been made to unify the Maron/Kuhns model with the Robertson/Sparck Jones model (Robertson, Maron and Cooper 1982). As indicated in Section 8, this unification suggests the possibility of using relevance feedback both locally (for the immediate query) and globally (to modify the document indexing for subsequent queries). Again, this model has not been tested experimentally, although some other techniques directed at the same end have been (see the discussion of Fuhr's work below).

9.3 Dependency

Following the original Robertson/Sparck Jones model (with its assumptions of independence of terms), a substantial amount of work was done (e.g. by van Rijsbergen and colleagues, see Harper and van Rijsbergen (1978)) on formal models which made some attempt to avoid or relax such assumptions.

These models were tested to some extent at the time they were developed (with the test collections available the time, which were extremely limited compared to the present generation of TREC-derived material). But the dependence models did not lead to any substantial improvements in performance.

There has been no substantial more recent work on dependence models supported by serious retrieval experiments, and indeed the practical challenges of computing dependencies on the large scale are very considerable. However much of the work done within the TREC Programme on the use of phrases and passages, for instance, can be seen as seeking to capture dependencies by more informal means, though there may be other motivations as well. Thus limiting candidate query expansion terms to those occurring in the passage neighbourhoods of matching terms can be seen as a way of concentrating co-occurrence information so that it is more discriminating than co-occurrences computed over extended full texts would be. Such techniques, as illustrated for example by the Local Context Analysis used with INQUERY (Xu and Croft 1996), have become quite common (Sparck Jones ()sparck98b,sparck98c), being taken as contributing, if only modestly, to performance.

9.4 Logical information retrieval

More recent work by van Rijsbergen has been in the area of logic and information retrieval, but with a particular probabilistic view incorporated into the logic (van Rijsbergen 1986, Sebastiani 1998). The essence of this approach is to re-interpret the basic concept of relevance as a logical relation between document and query (a document is relevant if it "implies" the query, in a way analogous to theorem-proving). It is then assumed that the information available is in some sense incomplete, so that the implication cannot generally be proved without the addition of missing information. The measure of this missing information is then the probability of relevance.

This work has stimulated a great deal of theoretical discussion, but generally little experimentation, and this only with drastic simplification. The exception is the work of Turtle and Croft and colleagues on the INQUERY system, discussed further below.

9.5 Networks

One class of probabilistic models which has been used extensively in other application areas consists of those based on networks. Such models see the domain of application as a network of nodes, with probabilistic relations between them. Generally the links are taken to represent the important or significant relationships. The absence of a direct link between two nodes is equated either with the absence of any relationship between them, or with the idea that any such relationship is a secondary one, implied by whatever multistep paths may exist between them.

In the retrieval area, there are several reasons to consider such models. We have plenty of candidate nodes, such as terms, documents and queries, and relationships between them that might be interpreted probabilistically. Indeed, in this paper we have several times appealed to arguments similar to those in the previous paragraph. For example, in the discussion on term frequencies (Section 4.5), we have supposed the existence of a hidden concept (node) associated with each term (the "eliteness" property), linked to some documents definitely but unknowably, and to the actual occurrences of the terms probabilistically. Furthermore the relation between term occurrences and relevance was assumed to be via the eliteness property, not directly.

Several authors have developed explicit network models for retrieval. The PIRCS system of Kwok is one such (Kwok 1995). This incorporates a "spreading activation" mechanism common to such models, whereby individual nodes are stimulated, and the stimulus spreads through the network via the links. In this case, retrieval involves stimulating a query node and allowing the stimulus to spread via terms to the documents. The most highly stimulated documents are then retrieved.

One component of this mechanism relates closely to the use of blind expansion, discussed in Section 5.3. PIRCS can be set up so that the stimulus passes from query to query terms, to documents indexed by those terms, back to the terms indexing those documents, and back to documents again. This is essentially the same mechanism as blind expansion without relevance information – the documents stimulated first are the initially retrieved items, and the terms stimulated from these documents are the equivalent of our expansion terms.

The INQUERY system of Croft and others (Turtle and Croft ()turtcroft90,turtcroft91) is also based on a network approach (as indicated above, their interpretation of probability of relevance is associated with the logical approach to retrieval). In comparison with the model presented in this paper, the inference network model on which INQUERY is based is potentially richer, because additional nodes and links may be incorporated. For example, apart from single query terms, if the query is expressed in Boolean form, the various Boolean constructions can be represented as nodes in their own right. Evidence from different nodes may be combined in different ways: in the Boolean case, this combination can be made to fit the normal Boolean logic (or extensions of it); or it can follow the sum-of-weights methods used in this paper. Other evidence-combination methods may also be defined.

Both PIRCS and INQUERY have been extensively tested in the TREC Programme; both tend to do well there. Again, there is a multitude of differences between Okapi and either PIRCS or INQUERY, both in matters of principle and in details; but there is also some commonality, and some mutual learning over successive rounds of TREC. As a result it is difficult to attribute the relatively small differences of performance to specific causes.

9.6 Regression

Some use has been made of regression models in IR (Cooper, Chen and Gey 1994)see e.g.. Essentially, a regression model takes the form of an assumed relationship between a dependent variable (relevance) and any independent variables which might be significant predictors of the dependent variable. The model then provides methods of estimating the parameters of the model directly from training data.

The probabilistic approach is in some sense complementary to the regression approach, in that regression could be and indeed has been used without any reference to the probabilistic model just as the probabilistic model can be used independent of regression (as in this paper). However probabilistic ideas have informed the regression approach in a number of ways. The first is that the dependent variable is generally taken to be *probability* of relevance rather than relevance itself. Second, because of the nature of relevance (assumed binary) and of probability, researchers investigating regression have tended to use logistic regression rather than traditional linear or polynomial regression. Third, some of the forms of relationships assumed in the regression models have been based on those found in probabilistic models (although in this context, regression is eclectic – any form of relationship might be regarded as candidate for a regression approach, whether the form is derived from a model, or observed empirically, or arrived at by any other means).

Despite the previous statement, only a rather limited range of relationships appear to be suitable for regression in practice. For example, it seems that the equations and parameters which are suggested by the probabilistic model (e.g. those involving k_1 or b) are not in a suitable form for learning by regression. On the whole, regression methods have been used to learn about general characteristics of terms. Fuhr's group have applied them both to searching and to the indexing stage (Fuhr and Buckley 1991, Fuhr et al. 1994)see e.g..

Work in this area has had mixed success. One problem is that the optimisation criterion for regression (something like least-squares error) is not necessarily well related to the retrieval performance measures (e.g. Average Precision) by which it is judged. Nevertheless, the Berkeley group (Cooper, Chen and Gey 1994) has had some success at TREC with a version of logistic regression used only when searching.

One observation to emerge from this Berkeley group's work is that if a log-odds score is converted back to an estimate of probability of relevance, documents at the top of the ranking often appear to have absurdly high probabilities. This presumably reflects the inaccuracy of the independence assumptions: combinations of terms do not really imply the kind of overwhelming evidence of relevance that the independence assumptions would suggest. This observation may result in a correction element being applied to the scoring method, which may, depending on its exact form, have little or no effect on the ranking of the documents, but would provide a more accurate estimate of the actual probability of relevance. Such a correction may be of value in the filtering task (see Section 7).

9.7 Other models

Very recently, a Hidden Markov Model approach familiar from so-called language modelling in speech recognition has been applied to information retrieval, with promising results (see e.g. Hiemstra ()hiemstra00, Miller, Leek and Schwartz ()bbn99; a related approach is taken by Ponte and Croft ()pontecroft98). It is, however, too soon to attempt a full assessment of its relation to our model.

10 Assessment

The object of this section is three-fold. First, we present a summary and overview of the results we have reported, identifying the main conclusions to be drawn about particular strategies and techniques for information retrieval, within the context of the probabilistic model as presented here. Second, we make some general remarks about this probabilistic model and its role in IR theory. Finally, we discuss a range of open issues around the boundaries of the work presented here.

10.1 Test summary and review

We have presented the development of the probabilistic model in successive steps, with accompanying test results. The latter were selected to cover specific points, e.g. whether some particular type of information, used in some particular model-defined way, was advantageous compared with not using this information, in this way. The individual test results were those given in Table 6, and the sequence of comparisons we have made was listed in Table 8, both in the Appendix to Part 1. It is now necessary to consider our set of experiments as a whole, in two ways: first, to take a broader view of the relative, and overall impact on performance of the various data types and retrieval strategies or more specific devices we have detailed; and second to consider the effect of different collection characteristics. We have hitherto been concerned to establish that strategy differences hold across collections and, for our TREC collection, across request forms. However some strategies may be relatively more advantageous under some collection conditions than others, with respect to request or document properties.

As a starting point we give the actual performance figures and significance test data for key runs and comparisons for the TREC collection, covering the major retrieval strategies in exemplar particular instantiations, in Table 14.

Then the overall outcome of our whole series of tests, across data types and strategies, can be summarised as follows. For the older collections, as reported in earlier publications, CFW gives some gain over UW, and RW over CFW. Absolute performance is quite good for RW, even with little relevance information. The same relations hold for TREC, except for the Very short requests, but absolute performance is very low.

TREC is the only collection to which further options apply. So we now focus on it and, considering both Doc30 and Rec30, look for *large* performance differences, which we define here as ones that are at least Noticeable but are typically more than that and hence are likely to be practically useful. This gives the overview shown in Table 15. In the table strategies are grouped, in an informal but intuitive way, into major and subordinate ones, and we also show points about yardsticks.

Thus looking first at the major strategies when we go beyond RW, it is evident that using within-document term frequencies is very valuable, but that using relevance information as well only comes into its own when it is used not for weighting alone, but for expansion: though even here the gain is not very large, or quite complete. In making these summary assessments we are deliberately suppressing performance detail, but that given earlier supports this broad brush picture with the required significance test checking as well as a range of more specific comparisons.

With the subordinate strategies on the other hand, there are no large gains to be made by what we may call elaboration, for instance exploiting passage-level matching, giving some extra weight to initial query terms in expansion, or fussing about the precise degree of expansion or trying to push expansion beyond a moderate level.

The yardstick use of relevance information, on the other hand, confirms the value of this information, subject to the qualification that has to be made about its application with the expansion strategy.

Turning now to absolute performance for the TREC case, and taking $QACIW + E$ (with

Table 14: Key runs and comparisons, TREC T741000X H collection

(figures rounded)

	Doc30			Rec30			AveP		
	L	M	V	L	M	V	L	M	V
UW	.04	.09	.15	.01	.05	.13	.01	.04	.09
CFW	.07	.15	.17	.04	.10	.17	.03	.07	.12
RW pred rel in 10	.21	.23	.18	.17	.20	.18	.12	.14	.12
CW	.41	.40	.40	.32	.32	.34	.23	.23	.24
QACW	.50	.44	.40	.44	.37	.34	.32	.27	.24
QACIW pred rel in 10	.52	.46	.42	.46	.42	.36	.33	.28	.25
QACIW + E pred rel in 10	.57	.52	.43	.51	.47	.36	.37	.34	.26

(exp L=32, M =24, V=16)

Significance tests, Wilcoxon

s = significant at the 1% level
m = significant at the 2.5% level
x = not significant

CFW	vs	UW	s	s	s	s	s	s	s	s	s
RW		CFW	s	s	s	s	s	s	s	s	s
CW		CFW	s	s	s	s	s	s	s	s	s
QACW		CFW *									
QACIW		QACW	s	s	s	s	m	s	s	s	s
QACIW + E		QACIW	s	s	s	s	s	s	x	x	x

*not run, use CW vs CFW as given

default expansion) as the best strategy along with the use of rel in 10 as a realistic application of relevance information, we find that Precision at Doc30 is .57 for the best case, with the Long requests, with a corresponding value of .51 for Rec30. Average Precision is .37. This is a very good level of performance.

With respect to the internal differences between request forms for TREC, the main point to note is that expansion is of more value for the Medium requests than the other two: presumably the Long forms do not need it, while it cannot be properly directed from the Very short ones. But more generally, gains from the various strategies explored are least for the Very short requests and are sometimes not Noticeable, though they are for the other forms. This is important because such brief requests are most likely to be encountered in practice. But it is also important that there are gains from the strategies (and hence the model) even for this kind of request.

On the absolute level of performance, taking the same $QACIW + E$ rel in 10 case as previously, we find that the difference between the request forms ranges from .57 for the Long requests, through .52 for Medium to .44 for Very short, with corresponding values for Rec30

Table 15: Overview of results, TREC T741000X collection

all request forms	
large performance differences at least Noticeable, typically more	
	exceptions
1) major strategies	
CFW large gain on UW	
RW large gain on CFW	V
(QA) CW very large gain on CFW	
QACIW very large gain on RW	
QACIW no gain on QACW if little rel info	
QACIW + E large gain on QACIW even if little rel info	V
QACIW + E no gain on QACIW with blind ‘rel’	
2) subordinate strategies	
QACIW + E heavy expansion no gain	
QACIW + E query term emphasis no gain	
QACIW, QACIW + E passage matching no gain	
3) yardsticks	
RW retro large gain on pred if pred little rel	
QACIW retro large gain on pred if pred little rel	
QACIW + E retro large gain on pred if pred little rel	L

of .51, .47 and .37. Average Precision is respectively .37, .34, and .27. Performance for the V requests, though much less impressive than for the Long ones, is still adequate.

Finally, we can assess the merit of the model via the strategies it implies. The largest single all round gains are made by using term frequency information, which is not very original or exciting. But it is also the case that when the best use of relevance information is *directly* compared with not using any, i.e. $QACIW + E$ with $QACW$, there is a further Noticeable gain in performance even for the V requests.

It is normally assumed that using full text compared with only titles+abstracts, or titles alone, on the document side, helps retrieval even with brief requests, but there is no strong evidence for this. So it is unfortunate that the collections we had did not permit systematic comparisons between different request forms for other than full text documents.

Essentially, what all the tests taken together endorse is a conclusion about effective retrieval strategies that seems weak but is in fact much stronger than it appears. From one point of view it is unfortunate that many plausible and widely applied devices, like blind relevance feedback, do not contribute much to performance, so that automatic retrieval systems are obliged to cover themselves as best they can in the flimsy garments supplied by term frequencies. The much more positive, opposite view is that the basic weighting methods supply the superb foundation figure on which all the rest of the wardrobe may be more or less seductively draped.

10.2 The model's status in IR theory

We have shown how a simple probabilistic model motivates a range of strategies and tactics for the use of certain categories of information within specifc weighting and scoring formulae to be used in retrieval. The experimental results presented have both confirmed the power of these strategies and tactics, and provided suitable values for certain tuning constants that occur in the model.

The probabilistic model is clearly not the only way to approach these issues. Many of our formulae are like (in behaviour if not in form) other formulae, motivated by other theories and/or by pragmatic considerations, that been successfully used in IR. We certainly cannot claim any unique validity or power for the probabilistic model, or for the specific formulae we have presented. However, we believe we have shown that the model provides a good and comprehensible basis for a systematic exposition of the components of the formula and their interrelations.

It is also the case that information retrieval does not depend exclusively on formulae. We have taken a very simple-minded view of the linguistic, semantic and epistemological issues involved. We rely on the fact, which is very apparent in text retrieval, that the language (English in particular, but not exclusively) allows us to identify content-bearing units relatively easily, and with little concern for the finer issues. Furthermore, these content-bearing units are extremely rich from the point of view of their use in retrieval. This does not, of course, preclude the possibility that more sophisticated approaches in any or all of these areas may provide even richer descriptions.

Within the scope of the probabilistic approach to IR, there are in fact many different (and not always compatible) ideas, concerning both the basic formulation of the model and its development; some of these ideas were discussed in Section 9, Comparisons. While the Probability Ranking Principle (Section 2, Part 1, and Robertson ()roberts77) is of very general applicability, the model as further developed and presented here cannot be regarded as having an exclusive claim on a justification in probability theory. However, at least part of the usefulness of the present model is that it translates directly into a retrieval mechanism based on a simple query-document matching or scoring function.

All in all, we suggest that the probabilistic model described in this paper is both reasonably well-founded and of clear and substantial value in the design of information retrieval systems.

10.3 Open issues

There are nevertheless important open issues to address in future: these are discussed in more detail in TR446 (1998), so are only summarised here. One is query length. Thus though we have used what we call "very short" queries in the tests reported here, many Web searchers appear to submit as few as two terms or even just one. These would not provide a good enough starting platform for the techniques we have described, for example through failing to retrieve any relevant documents at high rank. Much more work is needed both on understanding the functionality of Web queries and searching and on how the probabilistic approach can be leveraged in such situations. The same applies to cross-language retrieval when the initial source-language query is translationally under-specified, so it is hard to focus on target-language relevant documents.

Again, while the model applies in principle to other types of document descriptor than simple natural language terms, some key types may imply very short descriptions with few development hooks, and there is a more general need to accommodate searching on several key types at once (as INQUERY Turtle and Croft ()turtcroft90 is designed to do), and indeed to consider how the probabilistic approach may be applied when document, and hence description, structure is regarded as important.

Where more than one type of index key applies, it may seem appropriate to use one type as a filter, in Boolean fashion, and then rank the filtered set as usual. But such two-stage processing does not naturally fit the model, and more generally when working with only a small document set the model assumptions may not hold: thus it may be improper to treat all documents not know to be relevant as non-relevant, as in Section 4.2, Part 1. Similarly, in the same section it was asserted that the traditional *CFW* and the Croft/Harper (1979) approximation based on *RW* were very similar, but this would not hold if a term occurred in over half the document set, as could well apply to a selected subset. There are a variety of situations where two-stage searching could be rational, so these points need attention. This also, in a general way, applies to extending the model to other tasks, as discussed in Section 7, for example applying it to single document texts.

Finally, the practical efficacy of the model for text retrieval depends on the fact that that words are natural content-bearing units. Extending it to, for instance, other media, raises the question of whether such natural content units can be readily identified. The probabilistic model is a powerful apparatus for tackling 'rough' information management tasks: as Sparck Jones (1999) suggests, seeing just how far it can be pushed is an exciting challenge.

Acknowledgements

We are most grateful to our referees for their very helpful comments.

References

Biebricher, B., Fuhr, N., Lustig, G. Schwanter, M. and Knorz, G. (1988) The automatic indexing system AIR/PHYS – from research to application. *Proceedings of the 11th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval* New York: Association for Computing Machinery, 333–342.

Buckley, C., Allan, J. and Salton, G. (1995) Automatic retrieval and routing using SMART: TREC–2. *Information Processing and Management*, 31, 315–326.

Buckley, C., Singhal, A. and Mitra, M. (1996) New retrieval approaches using SMART: TREC–4. *The Fourth Text REtrieval Conference (TREC–4)*, (Ed. D.K. Harman) Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, 25–48.

Callan, J.P., Croft, W.B. and Broglio, J. (1995) TREC and TIPSTER experiments with INQUERY. *Information Processing and Management*, 31, 327–343.

Cooper, W. (1995) Some inconsistencies and misidentified modelling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13, 100–111.

Cooper, W., Chen, A. and Gey, F. (1994) Full text retrieval based on probabilistic equaitions with coefficients fitted by logistic regression. *The Second Text REtrieval Conference*

(TREC–4), (Ed. D.K. Harman), Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg, MD, 57–66.

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

Evans, D.A, and Lefferts, R.G. (1995) CLARIT–TREC experiments. *Information Processing and Management*, 31, 385–395.

Fuhr, N. and Buckley, C. (1991) A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9, 223–248.

Fuhr, N., Pfeifer, U. Bremkamp, C, and Pollman, M. (1994) Probabilistic learning approaches for indexing and retrieval with the TREC–2 collection. *The Second Text REtrieval Conference (TREC–4)*, (Ed. D.K. Harman), Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg, MD, 1997, 67–74.

Hand, T.F. (1997) A proposal for task-based evaluation of text summarisation systems. In *Intelligent, scaleable text summarisation* (Ed. I. Mani and M. Maybury), Proceedings of a Workshop, Somerset, NJ: Association for Computational Linguistics, 1997, 31–38.

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. Parts 1 and 2. *Journal of the American Society for Information Science*, 26, 197–206 and 280–289.

Hawking, D. and Thistlethwaite, P. (1998) Overview of TREC–6 very large collection track. *The Sixth Text REtrieval Conference (TREC–6)*, Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, 1998, 93–105.

Hiemstra, D. (2000) A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3 (to appear).

Ide, E. (1968) New experiments in relevance feedback. In *Scientific Report ISR–14*, Cornell University. Reprinted as Chapter 16 in *The SMART retrieval system* (Ed. Salton), Englewood Cliffs, NJ: Prentice-Hall, 1971.

Johnson, S.E., Jourlin, P., Sparck Jones, K. and Woodland, P.C. (1999) Spoken document retrieval for TREC–8 at Cambridge University. To appear in *The Eighth Text REtrieval Conference (TREC–8).*

Jones, G.J.F., Foote, J.T., Sparck Jones, K. and Young, S.J. (1996) Retrieving spoken documents by combining multiple evidence sources. *SIGIR 96: Proceedings of the 19th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 30–38.

Kwok, K.L. (1995) A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*, 13, 325–353.

Lewis, D. (1997) The TREC–5 filtering track. In *The Fifth Text REtrieval Conference (TREC–5)* (Ed. E.M. Voorhees and D.K. Harman), Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, 75–96.

Mani, I. and Maybury, M. (Eds.) (1997) *Intelligent, scaleable text summarisation*, Proceedings of a Workshop, Somerset, NJ: Association for Computational Linguistics, 1997.

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

Miller, D.R.H., Leek, T. and Schwartz, R.M. (1999) A Hidden Markov Model information retrieval system. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: Association for Computing Machinery, 214–221.

Mitra, M., Buckley, C., Singhal, A. and Cardie, C. (1997) An analysis of statistical and syntactic phrases. *Proceedings of RIAO–97, Computer-Assisted Information Searching on Internet*, Centre de Hautes Etudes Internationales d'Informatique Documentaires, Paris.

Okapi (1997) Papers on Okapi, Special Issue of *Journal of Documentation*, 33, 3–87.

Ponte, J. and Croft, W.B. (1998) A language modelling approach to information retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: Association for Computing Machinery, 275–281.

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

van Rijsbergen, C.J. (1979) *Information retrieval.* 2nd Ed, London: Butterworths.

van Rijsbergen, C.J. (1986) A non-classical logic for information retrieval. *The Computer Journal*, 29, 481–485.

Robertson, S.E. (1977) The probability ranking principle in IR. *Journal of Documentation*, 33, 294–304.

Robertson, S.E. (1990) On term selection for query expansion. *Journal of Documentation*, 46, 359–364.

Robertson, S.E. and Bovey, J.D. (1982) Statistical problems in the application of probabilistic models to information retrieval. Technical Report, Centre for Information Science, City University.

Robertson, S.E., Maron, and Cooper, W.S. (1982) Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1, 1–21.

Robertson, S.E., van Rijsbergen, C.J. and Porter, M.F. (1981) Probabilistic models of indexing and searching. In *Information retrieval research* (Ed. W.R. Oddy et al.). London: Butterworths, 35–65.

Robertson, S.E. and Walker, S. (1994) Some simple effective approximations to the 2 Poisson model for probabilistic weighted retrieval. *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 232–241.

Robertson, S.E. and Walker, S. (2000) Threshold setting in adaptive filtering. *Journal of Documentation*, 56, in press.

Rocchio, J.J. (1965) Relevance feedback in information retrieval. In *Scientific Report ISR–9*, Harvard University. Reprinted as Chapter 14 in *The SMART retrieval system* (Ed. G. Salton). Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

Salton, G. and Buckley, C. (1988) Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24, 513–523.

Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval*. Englewood Cliffs, NJ: Prentice Hall.

Salton, G., Singhal, A. Mitra, M. and Buckley, C. (1997) Automatic text structuring and summarisation. *Information Processing and Management*, 33, 193–207.

Sebastiani, F. (1998) On the role of logic in information retrieval. *Information Processing and Management*, 38 (1), 1–18.

Singhal, A., Buckley, C. and Mitra, M. (1996) Pivoted document length normalisation. *Proceedings of the 19th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval* New York, Association for Computing Machinery, 11–29.

Smeaton, A. and Wilkinson, R. (1997) Spanish and Chinese document retrieval in TREC–5. In *Proceedings of the Fifth Text REtrieval Conference (TREC–5)* (Ed. E.M. Voorhees and D.K. Harman), Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, 57–64.

Sparck Jones, K. (1999) (1999a) What is the role of NLP in text retrieval? In *Natural language information retrieval* (Ed. T. Strzalkowski), Dordrecht: Kluwer, 1–24.

Sparck Jones, K. (1999) (1999b) Summary performance comparisons: TREC–2 through TREC–7. In *The Seventh Text REtrieval Conference (TREC–7)*, Special Publication 500-242, National Institute of Standards and Technology, Gaithersburg, MD, B1–B6.

Sparck Jones, K. (1999) (1999c) Further reflections on TREC. *Information Processing and Management*, in press.

Sparck Jones, K. (1999) (1999d) Information retrieval and artificial intelligence. *Artificial Intelligence*, in press.

Sparck Jones, K., Jones, G.J.F., Foote, J.T. and Young. S.J. (1996) Experiments in spoken document retrieval. *Information Processing and Management*, 32, 399–419.

Sparck Jones, K., Walker, S. and Robertson, S.E. (1998) A probabilistic model of information retrieval: development and status. TR 446, Computer Laboratory, University of Cambridge (via http://www.cl.cam.ac.uk/).

Sparck Jones, K. and Webster, C.A. (1980) Research on relevance weighting 1976–1979, Computer Laboratory, University of Cambridge (also BL R&D Report 5553).

Sparck Jones, K. and Willett, P. (Eds.) (1997) *Readings in information retrieval*, San Francisco: Morgan Kaufmann.

TREC (1992–1999): D.K. Harman (Ed.) *The First Text REtrieval Conference (TREC–1)*, Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, 1993; *... Second ... (TREC–2)*. SP 500-215, 1994; *... Third ...(TREC–3)*, SP 500-225, 1995; *... Fourth ... (TREC–4)*, SP 500-236, 1996; Voorhees, E.M. and Harman, D.K. (Eds.) *... Fifth ... (TREC–5)*, SP 500-238, 1997; *... Sixth ... (TREC–6)*, SP 500-240, 1998; *... Seventh ... (TREC–7)*, SP 500-242, 1999.

Turtle, H.R. and Croft, W.B. (1990) Inference networks for document retrieval. *Proceedings of the 13th International ACM–SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 1–24, 1990.

Turtle, H.R. and Croft W.B. (1991) Evaluation of an inference network-based retrieval model, *ACM Transactions on Information Systems*, 7, 187–222.

Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F., and Sparck Jones, K. Okapi at TREC–6: automatic ad hoc, VLC, routing, filtering and QSDR. *Proceedings of the Sixth Text REtrieval Conference (TREC–6)*, Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, 1998, 125–136.

Xu, J. and Croft, W.B. (1996) *Proceedings of the 19th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval* New York, Association for Computing Machinery, 4–11.