On relevance weights with little relevance information

S.E. Robertson and S. Walker

Centre for Interactive Systems Research
Department of Information Science
City University, Northampton Square
London EC1V 0HB, UK

Abstract

The relationship between the Robertson/Sparck Jones relevance weighting formula and the Croft/Harper version for no relevance information is discussed. A method of avoiding the negative weights sometimes implied by the Croft/Harper version is proposed, which turns out to involve a return to the original Sparck Jones inverse collection frequency weight. The paper then goes on to propose a new way of using small amounts of relevance information in the estimation of relevance weights. Some experiments using TREC data are reported.

1 A short history

This paper concerns the weighting of search terms as a searching mechanism, when there may be some relevance judgements on a few documents in relation to the query or information need for which the search is being made. In this section, some elements of the history of relevance weighting are reviewed, as background to the present paper. Only those aspects relevant to the present discussion are covered.

1.1 Inverse collection frequency weighting

Weighting terms according to the number of documents in which they occur or to which they are assigned was discovered empirically by Sparck Jones to be an effective search device [1]. Frequent terms (that is, those that occur in many documents) are not in general good discriminators, and should be given low weights; thus the weight should be inversely related to the frequency. Such weighting is known as inverse document frequency (IDF) or inverse collection frequency. The usual formula is

$$w = \log \frac{N}{n}$$

where w is the weight to be assigned to the term, N is the size of (number of documents in) the collection and n is the number of documents in which the term occurs.

1.2 Robertson/Sparck Jones relevance weights

If the user has provided some relevance judgements for the information need for which the search is being made, then we may make use of these judgements to help derive good term weights for subsequent searching. Robertson and Sparck Jones [2] provide both theoretical and empirical support for the following weighting function:

$$w = \log \frac{p(1-q)}{q(1-p)} \tag{1}$$

where p is the probability that a document contains the term, given that it is relevant, and q is the same probability, given that it is not relevant.

If we have R relevant documents of which r contain the term, and N total documents of which n contain the term, then taking the obvious estimates of p and q gives the following:

$$w = \log \frac{r(N-R-n+r)}{(R-r)(n-r)} \tag{2}$$

However, for reasons to do with the estimation of logistic functions from small samples, the following formula is preferred:

$$w = \log \frac{(r+0.5)(N-R-n+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{3}$$

One obvious effect of the 0.5s is to prevent the formula from giving infinite weights under some circumstances.

It may also be observed that any of the above formulae may be separated into a part that relates to relevant documents only, and another that relates to non-relevant documents only. For example, the basic probabilistic formula 1 can be expressed as

$$w = \log \frac{p}{(1-p)} - \log \frac{q}{(1-q)} \tag{4}$$

1.3 Croft/Harper argument

Croft and Harper [3] use the Robertson/Sparck Jones weighting formula to derive a weighting scheme for situations where there is no relevance information. They propose that in the absence of such evidence, we may assume that (for query terms at least) p takes a fixed value, and that (given that almost all documents in a collection of reasonable size are likely to be non-relevant) q may be estimated by

SIGIR 97 Philadelphia PA, USA
Copyright 1997 ACM 0-89791-836-3/97/7..$3.50

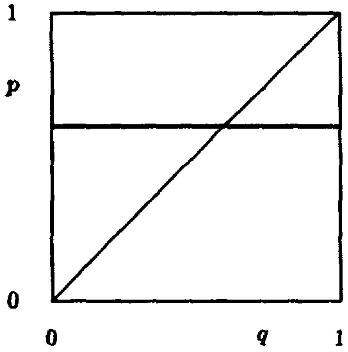
the proportion of items in the whole collection that contain the term. Then the formula 4 becomes

$$w = \text{Constant} + \log \frac{(N-n)}{n}$$

This formula is clearly closely related to the inverse collection frequency weight of section 1.1. In fact, if (as is usually the case) n is very much smaller than N, then the second part of the Croft/Harper formula is almost identical to the Sparck Jones inverse collection frequency weight. The constant in Croft/Harper is zero if $p = 0.5$, positive if $p > 0.5$, negative if $p < 0.5$.

The model used by Croft and Harper contains some anomalies. One is that $n = 0$ (or very small) and $p > 0$ are not really compatible, so that at this extreme the model seems to contain an inconsistency. But a more important anomaly (for the present paper) occurs at the other extreme, large n: here the formula predicts a negative weight for some terms. This can be seen graphically as follows. In Figure 1, showing a graph of p against q, the diagonal line ($p = q$) represents random terms, which are not correlated either positively or negatively with relevance and therefore have no value in retrieval. The horizontal line represents the Croft/Harper assumptions (constant p). The right-hand end of the line, below the diagonal, is that region where the Croft/Harper model gives negative weights: that is, it attributes to any terms in that region a negative correlation with relevance.

Figure 1: Croft/Harper assumptions



Since we are considering only query terms (*i.e.* terms used by the user to represent their information need), this effect is somewhat counter-intuitive. It is not normally a problem, because terms that occur in (*e.g.*) over half the collection are extremely rare.

1.4 The point-5 formula as an extension of Croft/Harper

The point-5 formula above can be seen as reducing to a simple version of the Croft/Harper model when there is no relevance information. If we take R as the number of *known* relevant documents, and therefore treat all unknown documents as non-relevant, then setting $R = r = 0$ (no relevance information) makes the point-5 formula 3 reduce to

$$w = \log \frac{N - n + 0.5}{n + 0.5}$$

which is virtually identical to the Croft/Harper formula with the constant set to zero (as discussed above, that is equivalent to setting $p = 0.5$).

If we now gain some relevance information, then we have some data from which to estimate p more precisely. Furthermore, as we discover documents to be relevant, we eliminate them from the assumed non-relevant set for estimating q. Thus the point-5 formula can be seen as providing a progressively better estimate of the weight, starting from a complete absence of relevance information, but responding to that information when it becomes available. This principle has been the basis of a number of relevance feedback systems, specifically the Okapi system [4].

There are several substantial *simplifications* (perhaps over-simplifications) in that argument. In particular, first, the initial assumption that $p = 0.5$ has been used without being properly tested — indeed, Croft and Harper's own experiments (on the Cranfield collection) suggested that a higher value of p would be appropriate. Second, the rate of response to relevance evidence might be badly out: it takes only a few relevant documents to swamp completely any remaining effect of the initial assumption. It might be better to rely on the initial assumption a little longer.

Third, the treatment of all documents not yet known to be relevant as non-relevant may be inappropriate: one limitation of this method is that it does not allow the relevance feedback system to take any account of explicit judgements of non-relevance. (However, it is worth also pointing out that Harper and van Rijsbergen [5] provide some evidence that it is better to do that than to rely entirely on such judgements of non-relevance for the estimation of q.)

2 A problem

The counter-intuitive negative weights referred to in section 1.3 would normally arise only in the case of a term which occurred in a very large proportion of the collection (over half if the value $p = 0.5$ is being assumed). As this is a very rare occurrence in most collections, this has not generally been seen as a problem.

However, we have recently come across a number of situations in which they could be a serious problem. Following are three such cases:

(a) We may wish to create a Boolean (or otherwise logical) limit set, within which the normal weighting-and-ranking methods operate. (The limit operation may represent certain properties which have to be present; this may occur for example in a database which combines well-defined numerical or logical DBMS fields with the sort of textual data common in information retrieval systems). In this case the collection size N would be replaced by the size of the limit set, which may be relatively small. Then the existence of some words which occur in over half of this limit set is not at all unlikely.

(b) We may have data that is logically nested, in the sense that one condition to which we want to assign a weight is logically implied by another such condition. An example might be a two-word phrase, when the single words are also present in the query. Here we may give the single words weights, and then assign a weight to the phrase which represents the extra value attached to the presence of the phrase, over and above that associated with the presence of both single words. In this case, it is appropriate to assign a weight to the phrase in relation to the size of the set defined by ANDing the two terms. Here again the phrase may easily occur in over half of the AND set.

(c) We may have a system which assigns weights to smaller units of language than the word. An example would be a Chinese language system which weights individual Chinese characters, or a voice retrieval system which weights phonemes. In both cases, it is entirely possible that some individual units occur in more than half the collection of documents.

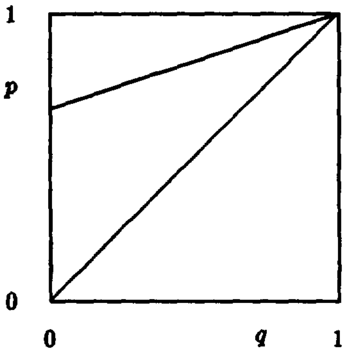
Associated with this problem of negative weights is the problem of how to deal with the non-relevance part of the weighting formula. As indicated in section 1.4, when we have some relevance information, we usually avoid taking specific account of judgements of non-relevance by treating all documents not known to be relevant as non-relevant. In the usual situation of a large collection, this makes some sense; but in any of the above situations, where we are weighting an entity with respect to a much smaller set, this approach becomes much less defendable.

It may be valuable to seek a weighting function which avoids both these problems.

3 An observation

If we return to the negative weight problem, and its cause as represented in Figure 1, we may consider alternative assumptions which may avoid the problem. An obvious one would be to assume that p, instead of remaining constant, increases from a non-zero starting point to reach unity with q. The simplest version of such a model would assume a straight line, as in Figure 2.

Figure 2: Alternative Croft/Harper assumptions



The straight-line model is actually rather intractable, and does not lead to a simple weighting formula. However, it is possible to construct a similar model, represented by a gentle curve which rises from a point on the p axis (actually 0.5) to (1,1) and which leads to a simple formula. The assumption of the model is that

$$q = \frac{n}{N}$$

as before (that is, with $R = r = 0$), while

$$p = \frac{1}{1 + \frac{N-n}{N}}$$

This leads to the weight

$$w = \log \frac{N}{n}$$

that is, the original Sparck Jones inverse collection frequency weight. This is an interesting result, in that it restores the original Sparck Jones formula to primary status, rather than suggesting that it is simply an approximation to the Croft/Harper version.

As in the Croft/Harper version, however, we can generalize to allow the starting-point of the line to be any point on the p axis, say p_0, with the following model:

$$p = \frac{p_0}{p_0 + (1-p_0)\frac{N-n}{N}}$$

which leads to the following weight:

$$w = \log \frac{p_0}{1-p_0} + \log \frac{N}{n}$$

The first part is of course a constant, as in the Croft/Harper version.

This generalization unfortunately may forfeit the property of avoiding negative weights. If $p_0 < 0.5$, the constant in the above equation is negative, the curve falls below the diagonal at the top end, and the weight is negative in that region. However, if $p_0 \geq 0.5$, the weight is always positive.

4 A variation

In this section, we propose a variation on the Robertson/Sparck Jones point-5 formula. The general principles of this proposal are those of the "rough model" suggested by Robertson and Walker [6]. That is, the general shape of the model — the relation between the weight and various other parameters — should follow that suggested by a probabilistic argument (as well as fitting with our empirical knowledge). However, there is no attempt to derive a complete probabilistic model from first principles — rather we seek a simple model which captures the general shape. Simplicity is partly a function of the formulae themselves, but mainly of the quantities that need to be estimated. We may expect the model not to be completely prescriptive, but to be tunable in the sense of having a small number of "constants" whose optimum values would be determined by experiment.

4.1 Principles

Following is a list of properties which, following the above analysis, we might demand of a relevance weighting function.

(a) With no relevance information, the function should give any term a reasonable prior weight.

(b) This prior weight should be (broadly speaking) an inverse collection frequency weight.

(c) It should not take negative values.

(d) It should be tunable.

(e) With a large amount of relevance information, the function should give an estimate which is entirely determined by the relevance evidence, and not at all related to the prior estimate.

(f) With a small amount of relevance information, the function should give a weight somewhere between the prior and a pure evidence-based estimate.

(g) The rate at which the estimate responds to new relevance information should be tunable.

(h) The estimates of the p and q components of the weight should be separate: for example, if we have only positive relevance judgements, the p component should be estimated from them, while the q component must be based on the prior.

The general approach taken in developing a relevance weighting function satisfying these conditions has been to estimate the weight (or rather its p and q components) directly, rather than via estimates for p and q themselves. This is consistent with the original justification for the point-5 formula and with some more recent work in probabilistic retrieval, such as the regression-based approaches [7]. Arguments based on estimating p and q are used initially, but then subsumed in estimates for the weight components.

4.2 Development

Basic formula

This is taken from 4 as

$$w = \log \frac{p}{(1-p)} - \log \frac{q}{(1-q)} = w_p - w_q$$

We seek estimating equations for w_p and w_q.

Prior for w_q

From the estimate $q = n/N$ used above, we can define a prior

$$w_q = \log \frac{n}{N-n}$$

Prior for w_p

From the estimate

$$p = \frac{p_0}{p_0 + (1-p_0)\frac{N-n}{N}}$$

used above, we can define a prior

$$w_p = k_4 + \log \frac{N}{N-n}$$

where $k_4 = \log \frac{p_0}{1-p_0}$

Prior weight

As above, these priors for w_p and w_q lead to a prior weight

$$w = k_4 + \log \frac{N}{n}$$

In order to satisfy condition (c) above, *i.e.* that the weight should not be negative, we need $k_4 \geq 0$, or $p_0 \geq 0.5$.

Evidence-based estimate for w_p

Given R relevant documents of which r contain the term, the pure evidence-based estimate should be

$$w_p = \log \frac{r+0.5}{R-r+0.5}$$

(The 0.5s are retained in this formula for the reason for which they were originally introduced into the Robertson/ Sparck Jones formula, because they minimise the bias in the estimate of the logistic function $\log(p/(1-p))$ for small samples. They are not now intended to deal with the no-relevance-information situation.)

The combination estimate should be a weighted average of the prior and evidence-based estimates, the weighting depending on the amount of evidence, that is on R. However, this combination should also be tunable. A possible way to achieve this is to combine the two components in the ratio $k : R$; this k is then the tuning constant. Then the combination becomes

$$w_p = \frac{k_5}{k_5+R}(k_4 + \log \frac{N}{N-n}) + \frac{R}{k_5+R} \log \frac{r+0.5}{R-r+0.5} \quad (5)$$

The basic assumption here, that the effect of the evidence-based estimate should be linear in R (which is the size of the evidence sample), is clearly questionable, but may be a suitable basis on which to start. A possible alternative assumption is that the effect should be linear in the square-root of R, on the grounds that the standard error of an estimate based on a sample is proportional to the square-root of the sample size. This would give an alternative to equation 5, as follows:

$$w_p = \frac{k_5}{k_5+\sqrt{R}}(k_4 + \log \frac{N}{N-n}) + \frac{\sqrt{R}}{k_5+\sqrt{R}} \log \frac{r+0.5}{R-r+0.5} \quad (6)$$

Evidence-based estimate for w_q

We are now dealing with known non-relevant items (as opposed to all documents not known to be relevant); hence we need to define some new notation. Let S be the number of known non-relevant items, and s be the number of these containing the term. The evidence-based estimate of w_q would then be

$$w_q = \log \frac{s+0.5}{S-s+0.5}$$

Again, we need an appropriate combination of this and the prior; on the same arguments as for w_p, we would have the following alternatives:

$$w_q = \frac{k_6}{k_6+S} \log \frac{n}{N-n} + \frac{S}{k_6+S} \log \frac{s+0.5}{S-s+0.5} \quad (7)$$

or

$$w_q = \frac{k_6}{k_6+\sqrt{S}} \log \frac{n}{N-n} + \frac{\sqrt{S}}{k_6+\sqrt{S}} \log \frac{s+0.5}{S-s+0.5} \quad (8)$$

(We may need a different tuning constant for these combinations than for the corresponding w_p combinations.)

Final combination weight

Combining our estimates for w_p and w_q, we get the following two possibilities: the linear combination (from equations 5 and 7),

$$\begin{aligned} w &= \frac{k_5}{k_5+R}(k_4 + \log\frac{N}{N-n}) + \frac{R}{k_5+R}\log\frac{r+0.5}{R-r+0.5} \\ &\quad -\frac{k_6}{k_6+S}\log\frac{n}{N-n} - \frac{S}{k_6+S}\log\frac{s+0.5}{S-s+0.5} \qquad (9) \end{aligned}$$

or the square-root combination (from equations 6 and 8):

$$\begin{aligned} w &= \frac{k_5}{k_5+\sqrt{R}}(k_4 + \log\frac{N}{N-n}) + \frac{\sqrt{R}}{k_5+\sqrt{R}}\log\frac{r+0.5}{R-r+0.5} \\ &\quad -\frac{k_6}{k_6+\sqrt{S}}\log\frac{n}{N-n} - \frac{\sqrt{S}}{k_6+\sqrt{S}}\log\frac{s+0.5}{S-s+0.5} \qquad (10) \end{aligned}$$

When $R = S = 0$, both these equations reduce to

$$w = k_4 + \log\frac{N}{n}$$

When R and S are large, they both approximate to a pure evidence-based weight:

$$w = \log\frac{(r+0.5)(S-s+0.5)}{(R-r+0.5)(s+0.5)}$$

(However, the approach to this approximation is slower in the case of equation 10 than for equation 9.)

In the experiments described below, where R and S are fixed for a particular run, equations 9 and 10 are equivalent with appropriate choice of k_5 and k_6. For example, equation 9 for $k_6 = x$ corresponds to equation 10 for $k_6 = x/\sqrt{S}$.

4.3 Discussion of tuning

k_4 essentially measures how good query terms are likely to be. The assumption built into the usual use of the point-5 formula is that $k_4 = 0$ (or $p_0 = 0.5$), and that would probably be a reasonable starting-point; but Croft and Harper's experiments suggested a somewhat larger value. Negative values are possible ($p_0 < 0.5$), but would re-introduce the problem of negative weights. Experiments to discover an appropriate value may be done without relevance information. It may be that terms from different sources (*e.g.* terms initially offered by the user *v.* terms offered by the system and selected by the user) have different optimum values of k_4.

k_5 and k_6 determine how quickly the estimate responds to evidence in the form of relevance judgements (respectively, positive or negative). Traditional use of the point-5 formula corresponds roughly (not precisely) to small k_5, say about 0.5, and very large k_6, that is to rapid response to positive relevance judgements but none at all to negative ones. Appropriate tuning experiments would involve using varying numbers of relevance judgements. These may be achieved by taking samples of available relevance judgements, or (eventually and more realistically) by using relevance judgements made on the top few ranked documents retrieved in an initial search.

5 Experimental results

A series of experiments have been undertaken on TREC data, along the lines suggested at the end of the previous section. These experiments are an initial exploration of the properties of the weighting functions proposed, and do not address the various situations mentioned in section 2.

5.1 Databases, queries and relevance judgments

Databases

The database for the retrospective searches was the TREC–5 routing database and the TREC–5 routing topics were used as the source of query terms. This database consists of 130,000 documents from the Foreign Broadcast Information Service. The mean document length is about 3400 characters (as indexed for the TREC–5 experiments).

For the predictive experiments a training database and two test databases were used. The training database used to weight the query terms consisted of alternate (even numbered) documents from TREC data disks 1, 2 and 3 with the addition of part of the data used for the TREC–4 routing experiments. This database contains 1,250,000 documents from various sources: newswires (AP, Wall Street Journal and San Jose Mercury News), Federal Register 1988, 1989 and 1994, Ziff (Articles from *Computer Select* disks), DOE abstracts, some US patents from 1993 and documents from Internet newsgroups. The mean document length is 2600 characters. The test databases were the TREC–5 routing database as used for the retrospective runs, and the other (odd numbered) half of the database of which the even half was used for training.

Queries

Queries were derived from the 39 TREC–5 routing topics which had six or more officially assessed relevant documents in the TREC–5 routing database. The TREC topic statements consist of a number of fields, always including a DESCRIPTION. Other fields, present in some of the topics, are TITLE, CONCEPTS and NARRATIVE. The topics used are not of a homogeneous nature; most (35) contain TITLE and NARRATIVE in addition to DESCRIPTION, and 27 of these also contain CONCEPTS. TREC results have shown that topics with CONCEPTs do better on the whole than those without this field. Three sets of query terms were derived from the topics: *short* queries from DESCRIPTIONs only, *medium* using additionally TITLEs and NARRATIVEs, and finally a *long* set which included CONCEPT terms whenever they were present.

Relevant fields from a typical TREC topic statement:

```
<num> Number: 011
<title> Topic: Space Program
<desc> Description:
Document discusses the goals or plans of the space
program  or a space project of any country or
organization.
<narr> Narrative:
To be relevant, a document must discuss the goals
or plans of a space program (e.g. the Space Station
Freedom) or space project (e.g. Shuttle mission
29-A) and identify the organization sponsoring the
program.

<con> Concept(s):

1. Shuttle, Space Plane, space station
2. Magellan, planetary explorer, satellites
3. vehicle launch
4. NASA, Ariane, European Space Agency (ESA)
5. Astronaut, Cosmonaut
6. Explorer, Dicsovery [sic], Columbia, Mir
7. Cape Canaveral, Star City
8. space
```

Topics were pre-processed to remove phrases like "to be relevant", then terms extracted from the required fields and 222 stop terms removed. Remaining terms then underwent a process of suffix-stripping (based on Porter [8]) and spelling normalization. The above topic gave the following query terms:

```
29a
europ
nasa
agenc
arian
astronaut
canaver
cape
citi
columbia
cosmonaut
countri
dicsoveri
eg
esa
explor
freedom
goal
launch
magellan
mir
mission
must
organ
plan
plane
planetar
program
project
satellit
shuttl
space
sponsor
star
station
vehicl
```

Finally, query terms were weighted using equation 10 with the desired values of R, S, k_4, k_5 and k_6. The mean query lengths (types per query) were long: 34.6, medium: 25.2 and short: 8.2. (Some trials were done in which account was taken of within-query term frequency, but these are not reported here.)

Relevance judgements

The official TREC relevance assessments were used. These are binary judgements, normally made by a single assessor for each topic, on the top-ranked documents retrieved by some of the participating systems. Although the relevance information is not complete in the way that it may be on a small test database, it may be assumed that nearly all the relevant documents are known, so nearly all the non-assessed documents are non-relevant. Thus, for each topic there are a number of known relevant documents, a (usually much larger) number of known non-relevant documents and a very large number of non-assessed documents each of which has only a very small probability of relevance.

For the purpose of these experiments N in equation 10 was the number of documents in the database used for weighting, and the 'R-' and 'S-' sets were taken from the assessed documents. The actual R and S sets for each experiment were systematically sampled from the documents assessed as relevant or non-relevant respectively, the sampling being designed to produce the desired number of documents.

For the test database the median number of relevant documents is 57 (mean 149, range 6–887); the corresponding figures for assessed non-relevant documents are 899 (918, 241–1398). For the training database the figures are 228 (250, 37–812); 1014 (1115, 474–2779).

Search and evaluation procedure

Apart from the new term-weighting functions and the use of relevance assessments all searches used the same Okapi software and general procedure as for City University's non-interactive TREC–4 [9] and TREC–5 [11] ad hoc runs, but with no account taken of within-query term frequency. That is, query terms were combined using the *BM25* function described in [9], the 1000 top-ranking document numbers were output and the average precision for each run calculated by means of the official TREC evaluation program.

5.2 The experiments

The primary object was to compare the effectiveness of the new weighting function (equation 10) with the original formula (equation 3), given various amounts of relevance information ranging from one or two assessed documents to, as a limiting case, a fully retrospective search on assumed complete relevance information.

A secondary experiment investigated the use of the new formula in a routing training situation, where relevance information from past assessments is used to derive query terms and weights for searching new documents. In effect, this differs from the first experiment in that searches do not retrieve any of the documents which have been used in "training".

5.3 Results

The scores reported in Tables 1–5 are average precisions ×1000. Most of the results reported are for the medium-length queries. The beneficial effect of both relevance and non-relevance information appeared to increase with query length, but the long queries are felt to be unrealistic and the results less worth reporting.

Perhaps the most striking result is that full use of the information from just a single relevant document gave a marked improvement in all cases. Negative information is much less useful than positive, but nevertheless gave a significant gain when used in conjunction with a reasonable amount of positive information.

Use of negative relevance judgements

It is clear from the tables that the negative relevance information is far less beneficial than the positive. For example, Table 1 shows that 10 known relevant documents alone increase the average precision of the searches by 61% for the medium queries. The corresponding gain for the short queries is 34%. However, in the retrospective experiment, non-zero S does give a significant further improvement, at least when $R \geq 5$; in the case $R = S = 10$ this was an additional 8% for the medium queries and 3% for the short queries. In a fully retrospective search, using all the positive and negative information gave total gains of 95% (medium queries) or 48% (short queries).

In the predictive experiments (tables 4 and 5) negative information was less useful, but appears still to be of some benefit.

Values for k_5 and k_6

Zero turned out to be the best value for k_5 in almost all cases. For k_6, if the linear formula 9 is used k_6 depends strongly on the value of S. But Tables 2 and 3 show that for the square root formula 10 k_6 values in the region of 4 to 16 work fairly well over quite a wide range of R and S.

Values for k_4

A number of experiments on k_4, without relevance information (i.e. with $R = S = 0$), indicated that a small negative value gave an improvement which was quite marked for the medium queries, less so for the short (tables 1 and 5). This reintroduces the problem of negative weights. However, as soon as there is any positive relevance information (even a

Topics were pre-processed to remove phrases like "to be relevant", then terms extracted from the required fields and 222 stop terms removed. Remaining terms then underwent a process of suffix-stripping (based on Porter [8]) and spelling normalization. The above topic gave the following query terms:

```
29a          dicsoveri    plan
europ        eg           plane
nasa         esa          planetar
agenc        explor       program
arian        freedom      project
astronaut    goal         satellit
canaver      launch       shuttl
cape         magellan     space
citi         mir          sponsor
columbia     mission      star
cosmonaut    must         station
countri      organ        vehicl
```

Finally, query terms were weighted using equation 10 with the desired values of R, S, k_4, k_5 and k_6. The mean query lengths (types per query) were long: 34.6, medium: 25.2 and short: 8.2. (Some trials were done in which account was taken of within-query term frequency, but these are not reported here.)

Relevance judgements

The official TREC relevance assessments were used. These are binary judgements, normally made by a single assessor for each topic, on the top-ranked documents retrieved by some of the participating systems. Although the relevance information is not complete in the way that it may be on a small test database, it may be assumed that nearly all the relevant documents are known, so nearly all the non-assessed documents are non-relevant. Thus, for each topic there are a number of known relevant documents, a (usually much larger) number of known non-relevant documents and a very large number of non-assessed documents each of which has only a very small probability of relevance.

For the purpose of these experiments N in equation 10 was the number of documents in the database used for weighting, and the 'R-' and 'S-' sets were taken from the assessed documents. The actual R and S sets for each experiment were systematically sampled from the documents assessed as relevant or non-relevant respectively, the sampling being designed to produce the desired number of documents.

For the test database the median number of relevant documents is 57 (mean 149, range 6–887); the corresponding figures for assessed non-relevant documents are 899 (918, 241–1398). For the training database the figures are 228 (250, 37–812); 1014 (1115, 474–2779).

Search and evaluation procedure

Apart from the new term-weighting functions and the use of relevance assessments all searches used the same Okapi software and general procedure as for City University's non-interactive TREC–4 [9] and TREC–5 [11] ad hoc runs, but with no account taken of within-query term frequency. That is, query terms were combined using the *BM25* function described in [9], the 1000 top-ranking document numbers were output and the average precision for each run calculated by means of the official TREC evaluation program.

5.2 The experiments

The primary object was to compare the effectiveness of the new weighting function (equation 10) with the original formula (equation 3), given various amounts of relevance information ranging from one or two assessed documents to, as a limiting case, a fully retrospective search on assumed complete relevance information.

A secondary experiment investigated the use of the new formula in a routing training situation, where relevance information from past assessments is used to derive query terms and weights for searching new documents. In effect, this differs from the first experiment in that searches do not retrieve any of the documents which have been used in "training".

5.3 Results

The scores reported in Tables 1–5 are average precisions ×1000. Most of the results reported are for the medium-length queries. The beneficial effect of both relevance and non-relevance information appeared to increase with query length, but the long queries are felt to be unrealistic and the results less worth reporting.

Perhaps the most striking result is that full use of the information from just a single relevant document gave a marked improvement in all cases. Negative information is much less useful than positive, but nevertheless gave a significant gain when used in conjunction with a reasonable amount of positive information.

Use of negative relevance judgements

It is clear from the tables that the negative relevance information is far less beneficial than the positive. For example, Table 1 shows that 10 known relevant documents alone increase the average precision of the searches by 61% for the medium queries. The corresponding gain for the short queries is 34%. However, in the retrospective experiment, non-zero S does give a significant further improvement, at least when $R \geq 5$; in the case $R = S = 10$ this was an additional 8% for the medium queries and 3% for the short queries. In a fully retrospective search, using all the positive and negative information gave total gains of 95% (medium queries) or 48% (short queries).

In the predictive experiments (tables 4 and 5) negative information was less useful, but appears still to be of some benefit.

Values for k_5 and k_6

Zero turned out to be the best value for k_5 in almost all cases. For k_6, if the linear formula 9 is used k_6 depends strongly on the value of S. But Tables 2 and 3 show that for the square root formula 10 k_6 values in the region of 4 to 16 work fairly well over quite a wide range of R and S.

Values for k_4

A number of experiments on k_4, without relevance information (i.e. with $R = S = 0$), indicated that a small negative value gave an improvement which was quite marked for the medium queries, less so for the short (tables 1 and 5). This reintroduces the problem of negative weights. However, as soon as there is any positive relevance information (even a

Table 1: Best scores for retrospective searches.

Figures are average precision × 1000. k_5 and k_6 were chosen so as to give the best result in each case.

	medium queries								
	S								
R	0	1	2	3	4	5	10	15	all
0	152	152	152	152	152	152	152	152	154
0	(171	$k_4 = -1$)							
1	201	203	202	204	202				209
2	235	239	238	241	242				236
3	233	236	235	237	235				264
4	261	266	267	266	267			269	260
5	265	271	269	269	271	270	274	273	292
10	275	293	292	291	293	291	298	299	302
15	286	305	306	304	307	308	312	316	314
all	295	315	316	318	319	317	326	326	334
	short queries								
	S								
R	0	1	2	3	4	5	10	15	all
0	153	153	153	153	153	153	153	153	153
0	(164	$k_4 = -1$)							
1	185	185	185	185	184				
2	201	201	200	201	201				
3	191	190	190	191	190				
4	204	204	204	204	204				
5	203	204	203	203	204	204	204	203	
10	220	223	222	224	223	224	227	224	
15	225	229	228	232	230	231	234	232	
all	231	235	238	238	241	237	241	240	243

Table 2: Effect of varying k_5 and k_6: retrospective searches, medium queries, square root formula 10

	k_6								
k_5	0	1	2	4	8	16	32	64	∞
	R =all, S =all								
0	248	265	280	298	319	334	333	322	295
1	254	267	278	301	321	327	323	307	280
2	245	262	277	296	311	317	310	296	268
4	235	249	262	279	295	300	292	278	252
	$R = S = 10$								
0	168	272	290	298	292	286	282	278	275
1	161	261	283	283	275	266	260	257	254
2	147	248	268	267	258	250	243	241	238
4	114	218	241	240	231	224	221	218	214
	$R = S = 5$								
0	124	225	256	268	270	269	267	266	265
1	114	221	250	251	255	247	245	242	240
2	104	208	234	235	229	225	222	221	220
4	088	182	204	208	203	200	199	198	198
	$R = S = 2$								
0	058	194	222	236	238	237	236	236	235
1	044	182	208	213	210	208	206	205	204
2	039	164	188	191	191	190	189	189	189
4	032	139	166	175	174	174	174	174	174

Table 3: Scores for retrospective searches at constant $k_5 = 0$ and $k_6 = 8$: medium queries, square root formula

	S								
R	0	1	2	3	4	5	10	15	all
0	152								
1	201	203	202	204	202				
2	235	239	238	241	242				
3	233	236	234	237	235				
4	261	265	266	266	267				
5	265	269	269	269	271	270	274	271	
10	275	283	285	286	288	287	292	297	
15	286	295	299	300	302	304	307	314	
all	295	303	309	310	314	310	319	321	319

Table 4: Best scores for predictive searches: TREC–5 routing database, medium queries

	S								
R	0	1	2	3	4	5	10	15	all
0	139	138	138	138	138	138	138	138	
1	175	175	175	175	175				
2	190	190	190	190	190				
3	211	211	211	212	211				
4	213	214	214	215	213	215	214		
5	208	209	208	209	208	209	208	208	
10	227	230	231	233	230	231	232	231	
15	230	232	231	233	232	234	231	232	
all	245	251	251	250	255	254	254	261	265

Table 5: Best scores for predictive searches: large database, medium queries

	S								
R	0	1	2	3	4	5	10	15	all
0	183								
0	(200	$k_4 = -1$)							
1	226	226	226	226	226				226
2	243	243	243	243	243				252
3	265	266	265	267	265				267
4	266	266	266	269	266	267			269
5	268	270	269	272	269	269	268	269	275
10	286	294	291	294	291	289	288	290	297
15	297	305	303	301	300	303	302	304	302
all	309	316	315	314	317	321	318	323	327

single known relevant document) the best results were obtained by using equation 10 with $k_5 = 0$, in which case k_4 has no effect.

6 Discussion and conclusions

It seems that there is some evidence that a small benefit could be obtained by including the constant k_4 in the no-relevance-information version of the inverse collection frequency weight. However, the benefit is not great, at least on the TREC data tested, and is obtained in a way which re-introduces the problem of negative weights.

The proposed method of combining prior and evidence-based estimates when some information is available is more promising. The results suggest that the simple linear combination, linear also in the sample sizes, may not be a very good one. However, the equation based on the square roots of the sample sizes, but still in the form of a linear combination of the prior and evidence-based estimates, looks much better.

It is also clear that in the case of p, concerning the relevant documents, explicit relevance evidence should take over from the prior quickly. In the case of q (non-relevant documents), explicit judgements of non-relevance are useful but should be allowed only a slow effect. It requires very much more such evidence before the prior should be thrown away.

It may be argued that the combination functions should impose a limit on the large-sample effect, in other words the prior should retain some influence even if the samples are large. An argument for such a limit might be made on the grounds that "samples" in relevance feedback are always biased (e.g. those documents that ranked highly in an earlier iteration or those that were retrieved from older material). Thus the searcher's initial choice of terms may be said to contain information which may not be reflected in the evidence. In the present experiments, this argument would not apply to the case of p, but might to q. When we progress to more realistic experiments, e.g. simulations of interactive searching or of a routing environment, such an argument may have more force.

We have yet to test the ideas in any of the environments suggested in section 2, where the negative weights of the Croft-Harper formula are more evidently problematic. Given the moderately promising results reported here, some work in the areas is suggested.

References

[1] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 1972; 28:11–21.

[2] Robertson S.E. and Sparck Jones K. Relevance weighting of search terms. *Journal of the American Society for Information Science* 1976; 27:129–146.

[3] Croft W. and Harper D. Using probabilistic models of information retrieval without relevance information. *Journal of Documentation* 1979; 35:285–295.

[4] Special issue on Okapi and information retrieval research. *Journal of Documentation* 1997; 53 no. 1.

[5] Harper D.J and van Rijsbergen C.J. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation* 1978; 34:189–216.

[6] Robertson S.E. and Walker S. Some simple effective approximations to the 2–Poisson model for probabilistic weighted retrieval. In: Croft W.B. and van Rijsbergen C.J. (eds) *SIGIR 94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994.* Springer-Verlag, 1994 (pp.232–241).

[7] Cooper W., Chen A. and Gey F. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In: Harman D. (ed.) *The Second Text Retrieval Conference (TREC–2)* (NIST Special Publication 500–215) NIST, 1994 (pp.57–66).

[8] Porter M F. An algorithm for suffix stripping. *Program 14* 3 Jul 1980 130-137.

[9] Robertson S E *et al.* Okapi at TREC–4. In: [10]. p73–96.

[10] *The fourth Text REtrieval Conference: TREC–4.* Edited by D.K. Harman. NIST, 1996.

[11] Beaulieu M M *et al.* Okapi at TREC–5. In: [12]. p?.

[12] [TREC–5 proceedings.] NIST. To be published.