

Some Simple Effective Approximations to the 2–Poisson Model for Probabilistic Weighted Retrieval

S.E. Robertson

ser@is.city.ac.uk

S. Walker

sw@is.city.ac.uk

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

Abstract

The 2–Poisson model for term frequencies is used to suggest ways of incorporating certain variables in probabilistic models for information retrieval. The variables concerned are within-document term frequency, document length, and within-query term frequency. Simple weighting functions are developed, and tested on the TREC test collection. Considerable performance improvements (over simple inverse collection frequency weighting) are demonstrated.

1 Introduction

This paper discusses an approach to the incorporation of new variables into traditional probabilistic models for information retrieval, and some experimental results relating thereto. Some of the discussion has appeared in the proceedings of the second TREC conference [1], albeit in less detail.

Statistical approaches to information retrieval have traditionally (to over-simplify grossly) taken two forms: firstly approaches based on formal models, where the model specifies an exact formula; and secondly ad-hoc approaches, where formulae are tried because they seem to be plausible. Both categories have had some notable successes. A more recent variant is the regression approach of Fuhr and Cooper (see, for example, Cooper [3]), which incorporates ad-hoc choice of independent variables and functions of them with a formal model for assessing their value in retrieval, selecting from among them and assigning weights to them.

One problem with the formal model approach is that it is often very difficult to take into account the wide variety of variables that are thought or known to influence retrieval. The difficulty arises either because there is no known basis for a model containing such variables, or because any such model may simply be too complex to give a usable exact formula.

One problem with the ad-hoc approach is that there is little guidance as to how to deal with specific variables—one has to guess at a formula and try it out. This problem is also apparent in the regression approach—although “trying it out” has a somewhat different sense here (the formula is tried in a regression model, rather than in a retrieval test).

The approach in the present paper is to take a model which provides an exact but intractable formula, and use it to suggest a much simpler formula. The simpler formula can then be tried in an ad-hoc fashion, or used in turn in a regression model. Although we have not yet taken this latter step of using regression, we believe that the present suggestion lends itself to such methods.

The variables we have included in this paper are: within-document term frequency, document length, and within-query term frequency (it is worth observing that collection frequency of terms appears naturally in traditional probabilistic models, particularly in the form of the approximation to inverse collection frequency weighting demonstrated by Croft and Harper [4]). The formal model which is used to investigate the effects of these variables is the 2–Poisson model (Harter [5], Robertson, van Rijsbergen and Porter [6]).

2 Basic Probabilistic Weighting Model

The basic weighting function used is that developed in [6], and may be expressed as follows:

$$w(\underline{\mathbf{x}}) = \log \frac{P(\underline{\mathbf{x}}|R) P(\mathbf{0}|\bar{R})}{P(\underline{\mathbf{x}}|\bar{R}) P(\mathbf{0}|R)}, \quad (1)$$

where $\underline{\mathbf{x}}$ is a vector of information about the document, $\mathbf{0}$ is a reference vector representing a zero-weighted document, and R and \bar{R} are relevance and non-relevance respectively.

For example, each component of $\underline{\mathbf{x}}$ may represent the presence/absence of a query term in the document or its document frequency; $\mathbf{0}$ would then be the “natural” zero vector representing all query terms absent.

In this formulation, independence assumptions (or, indeed, Cooper’s assumption of “linked dependence” [7]), lead to the decomposition of w into *additive* components such as individual term weights. In the presence/absence case, the resulting weighting function is the Robertson/Sparck Jones formula [8] for a term-presence-only weight, as follows:

$$w = \log \frac{p(1-q)}{q(1-p)}, \quad (2)$$

where $p = P(\text{term present}|R)$ and $q = P(\text{term present}|\bar{R})$.

With a suitable estimation method, this becomes:

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

where N is the number of indexed documents, n the number containing the term, R the number of known relevant documents, and r the number of these containing the term. This approximates to inverse collection frequency (ICF) when there is no relevance information. It will be referred to below (with or without relevance information) as $w^{(1)}$.

If we deal with within-document term frequencies rather than merely presence and absence of terms, then the formula corresponding to 2 would be as follows:

$$w = \log \frac{p_{tf}q_0}{q_{tf}p_0}, \quad (4)$$

where $p_{tf} = P(\text{term present with frequency } tf|R)$, q_{tf} is the corresponding probability for \bar{R} , and p_0 and q_0 are those for term absence.

2.1 Eliteness

The approach taken in reference [6] is to model within-document term frequencies by means of a mixture of two Poisson distributions. However, before discussing the 2-Poisson model, it is worth extracting one idea which is necessary for the model, but can perhaps stand on its own. (This idea was in fact taken from the original 2-Poisson work by Harter [5], but was extended somewhat to allow for multi-concept queries.)

We hypothesize that occurrences of a term in a document have a random or stochastic element, which nevertheless reflects a real but hidden distinction between those documents which are “about” the concept represented by the term and those which are not. Those documents which are “about” this concept are described as “elite” for the term. We may draw an inference about eliteness from the term frequency, but this inference will of course be probabilistic. Furthermore, relevance (to a query which may of course contain many concepts) is related to eliteness rather than directly to term frequency, which is assumed to depend *only* on eliteness. The usual term-independence assumptions are replaced by assumptions that the eliteness properties for different terms are independent of each other; although the details are not provided here, it is clear that the independence assumptions can be replaced by “linked dependence” assumptions in the style of Cooper [7].

As usual, the various assumptions of the model are very clearly over-simplifications of reality. It seems nevertheless to be useful to introduce this hidden variable of eliteness in order to gain some understanding of the relation between multiple term occurrence and relevance.

3 The 2-Poisson Model

The 2-Poisson model is a specific distributional assumption based on the eliteness hypothesis discussed above. The assumption is that the distribution of within-document frequencies is Poisson for the elite documents, and also (but with a different mean) for the non-elite documents.

It would be possible to derive this model from a more basic one, under which a document was a random stream of term occurrences, each one having a fixed, small probability of being the term in question, this probability being constant over all elite documents, and also constant (but smaller) over all non-elite documents. Such a model would require that all documents were the same length. Thus the 2-Poisson model is usually said to assume that document length is constant: although technically it does not require that assumption, it makes little sense without it. Document length is discussed further below (section 5).

The approach taken in [6] was to estimate the parameters of the two Poisson distributions for each term directly from the distribution of within-document frequencies. These parameters were then used in various weighting functions. However, little performance benefit was gained. This was seen essentially as a result of estimation problems: partly that the estimation method for the Poisson parameters was probably not very good, and partly because the model is complex in the sense of requiring a large number of different parameters to be estimated. Subsequent work on mixed-Poisson models has suggested that alternative estimation methods may be preferable [9].

Combining the 2-Poisson model with formula 4, under the various assumptions given about dependencies, we obtain [6] the following weight for a term t :

$$w = \log \frac{(p'\lambda^{tf}e^{-\lambda} + (1-p')\mu^{tf}e^{-\mu})(q'e^{-\lambda} + (1-q')e^{-\mu})}{(q'\lambda^{tf}e^{-\lambda} + (1-q')\mu^{tf}e^{-\mu})(p'e^{-\lambda} + (1-p')e^{-\mu})}, \quad (5)$$

where λ and μ are the Poisson means for tf in the elite and non-elite sets for t respectively, $p' = P(\text{document elite for } t|R)$, and q' is the corresponding probability for \bar{R} .

The estimation problem is very apparent from equation 5, in that there are four parameters for each term, for *none* of which are we likely to have direct evidence (because of eliteness being a hidden variable). It is precisely this estimation problem which makes the weighting function intractable. This consideration leads directly to the approach taken in the next section.

4 A Rough Model for Term Frequency

4.1 The Shape of the tf Effect

Many different functions have been used to allow within-document term frequency tf to influence the weight given to the particular document on account of the term in question. In some cases a linear function has been used; in others, the effect has been dampened by using a suitable transformation such as $\log tf$.

Even if we do not use the full equation 5, we may allow it to suggest the shape of an appropriate, but simpler, function. In fact, equation 5 has the following characteristics: (a) It is zero for $tf = 0$; (b) it increases monotonically with tf ; (c) but to an asymptotic maximum; (d) which approximates to the Robertson/Sparck Jones weight that would be given to a direct indicator of eliteness.

Only in an extreme case, where eliteness is identical to relevance, is the function linear in tf . These points can be seen from the following re-arrangement of equation 5:

$$w = \log \frac{(p' + (1-p')(\mu/\lambda)^{tf}e^{\lambda-\mu})(q'e^{\mu-\lambda} + (1-q'))}{(q' + (1-q')(\mu/\lambda)^{tf}e^{\lambda-\mu})(p'e^{\mu-\lambda} + (1-p'))}. \quad (6)$$

μ is smaller than λ . As $tf \rightarrow \infty$ (to give us the asymptotic maximum), $(\mu/\lambda)^{tf}$ goes to zero, so those components drop out. $e^{\mu-\lambda}$ will be small, so the approximation is:

$$w = \log \frac{p'(1-q')}{q'(1-p')}. \quad (7)$$

(The last approximation may not be a good one: for a poor and/or infrequent term, $e^{\mu-\lambda}$ will not be very small. Although this should not affect the component in the numerator, because q' is likely to be small, it will affect the component in the denominator.)

4.2 A Simple Formulation

What is required, therefore, is a simple tf -related weight that has something like the characteristics (a)-(d) listed in the previous section. Such a function can be constructed as follows. The function $tf/(\text{constant} + tf)$ increases from zero to an asymptotic maximum in approximately the right fashion. The constant determines the rate at which the increase drops off: with a large constant, the function

is approximately linear for small tf , whereas with a small constant, the effect of increasing tf rapidly diminishes.

This function has an asymptotic maximum of one, so it needs to be multiplied by an appropriate weight similar to equation 7. Since we cannot estimate 7 directly, the obvious simple alternative is the ordinary Robertson/Sparck Jones weight, equation 2, based on presence/absence of the term. Using the usual estimate of 2, namely $w^{(1)}$ (equation 3), we obtain the following weighting function:

$$w = \frac{tf}{(k_1 + tf)} w^{(1)}, \quad (8)$$

where k_1 is an unknown constant.

The model tells us nothing about what kind of value to expect for k_1 . Our approach has been to try out various values of k_1 (values around 1–2 seem to be about right for the TREC data—see the results section 7 below). However, in the longer term we hope to use regression methods to determine the constant. It is not, unfortunately, in a form directly susceptible to the methods of Fuhr or Cooper, but we hope to develop suitable methods.

The shape of formula 8 differs from that of formula 5 in one important respect: 8 is convex towards the upper left, whereas 5 can under some circumstances (that is, with some combinations of parameters) be S-shaped, increasing slowly at first, then more rapidly, then slowly again. Averaging over a number of terms with different values of the parameters is likely to reduce any such effect; however, it may be useful to try a function with this characteristic. One such, a simple combination of 8 with a logistic function, is as follows:

$$w = \frac{tf^c}{(k_1^c + tf^c)} w^{(1)}, \quad (9)$$

where $c > 1$ is another unknown constant. This function has not been tried in the present experiments.

5 Document Length

As indicated in section 3, the 2-Poisson model in effect assumes that documents (i.e. records) are all of equal length. Document length is a variable which figures in a number of weighting formulae.

5.1 Hypotheses Concerning Document Length

We may postulate at least two reasons why documents might vary in length. Some documents may simply cover more material than others; an extreme version of this hypothesis would have a long document consisting of a number of unrelated short documents concatenated together (the “Scope hypothesis”). An opposite view would have long documents like short documents, but longer: in other words, a long document covers a similar scope to a short document, but simply uses more words (the “Verbosity hypothesis”).

It seems likely that real document collections contain a mixture of these effects; individual long documents may be at either extreme or of some hybrid type. (It is worth observing that some of the long TREC news items read exactly as if they are made up of short items concatenated together.) However, with the exception of a short discussion in section 5.7, the work on document length reported in this paper assumes the Verbosity hypothesis; little progress has yet been made with models based on the Scope hypothesis.

The Verbosity hypothesis would imply that document properties such as relevance and eliteness can be regarded as independent of document length; given eliteness for a term, however, the number of occurrences of that term would depend on document length.

5.2 A Very Rough Model

The simplest way to incorporate this hypothesis is to take formula 8 above, but normalise tf for document length (d). If we assume that the value of k_1 is appropriate to documents of average length (Δ), then this model can be expressed as

$$w = \frac{tf}{\left(\frac{k_1 \times d}{\Delta} + tf\right)} w^{(1)}. \quad (10)$$

This function is used in the experiments described below (section 7). However, a more detailed analysis of the effect of the Verbosity hypothesis on the 2-Poisson model may reveal a more complex pattern.

5.3 Document Length in the Basic Model

Referring back to the basic weighting function 1, we may include document length as one component of the vector $\underline{\mathbf{x}}$. However, document length does not so obviously have a “natural” zero (an actual document of zero length is a pathological case). Instead, we may use the average length of a document for the corresponding component of the reference vector $\underline{\mathbf{Q}}$; thus we would expect to get a formula in which the document length component disappears for a document of average length, but not for other lengths. The weighting formula then becomes:

$$w(\underline{\mathbf{x}}, d) = \log \frac{P(\underline{\mathbf{x}}, d|R) P(\underline{\mathbf{Q}}, \Delta|\bar{R})}{P(\underline{\mathbf{x}}, d|\bar{R}) P(\underline{\mathbf{Q}}, \Delta|R)},$$

where d is document length, and $\underline{\mathbf{x}}$ represents all other information about the document. This may be decomposed into the sum of two components, $w(\underline{\mathbf{x}}, d)_1 + w(\underline{\mathbf{Q}}, d)_2$, where

$$w(\underline{\mathbf{x}}, d)_1 = \log \frac{P(\underline{\mathbf{x}}, d|R) P(\underline{\mathbf{Q}}, d|\bar{R})}{P(\underline{\mathbf{x}}, d|\bar{R}) P(\underline{\mathbf{Q}}, d|R)} \quad \text{and} \quad w(\underline{\mathbf{Q}}, d)_2 = \log \frac{P(\underline{\mathbf{Q}}, d|R) P(\underline{\mathbf{Q}}, \Delta|\bar{R})}{P(\underline{\mathbf{Q}}, d|\bar{R}) P(\underline{\mathbf{Q}}, \Delta|R)}. \quad (11)$$

These two components are discussed separately.

5.4 Consequences of the Verbosity Hypothesis

We assume without loss of generality that the two Poisson parameters for a given term, λ and μ , are appropriate for documents of average length. Then the Verbosity hypothesis would imply that while a longer (say) document has more words, each individual word has the same probability of being the term in question. Thus the distribution of term frequencies in documents of length d will be 2-Poisson with means $\lambda d/\Delta$ and $\mu d/\Delta$.

We may also make various independence assumptions, such as between document length and relevance.

5.5 Second Component

The second component of equation 11 is

$$w(\underline{\mathbf{Q}}, d)_2 = \log \frac{P(\underline{\mathbf{Q}}|R, d) P(\underline{\mathbf{Q}}|\bar{R}, \Delta)}{P(\underline{\mathbf{Q}}|\bar{R}, d) P(\underline{\mathbf{Q}}|R, \Delta)} + \log \frac{P(d|R) P(\Delta|\bar{R})}{P(d|\bar{R}) P(\Delta|R)}.$$

Under the Verbosity hypothesis, the second part of this formula is zero. Making the usual term-independence or linked-dependence assumptions, the first part may be decomposed into a sum of components for each query term, thus:

$$w(t, d)_2 = \log \frac{(p'e^{-\lambda d/\Delta} + (1-p')e^{-\mu d/\Delta}) (q'e^{-\lambda} + (1-q')e^{-\mu})}{(q'e^{-\lambda d/\Delta} + (1-q')e^{-\mu d/\Delta}) (p'e^{-\lambda} + (1-p')e^{-\mu})}. \quad (12)$$

Note that because we are using the zero-vector $\underline{\mathbf{Q}}$, there is a component for each query term, whether or not the term is in the document.

For almost all normal query terms (i.e. for any terms that are not actually detrimental to the query), we can assume that $p' > q'$ and $\lambda > \mu$. In this case, formula 12 can be shown to be monotonic decreasing with d , from a maximum as $d \rightarrow 0$, through zero when $d = \Delta$, and to a minimum as $d \rightarrow \infty$. As indicated, there is one such factor for each of the nq query terms.

Once again, we can devise a very much simpler function which approximates to this behaviour, as follows:

$$\text{correction factor} = k_2 \times nq \frac{(\Delta - d)}{(\Delta + d)}, \quad (13)$$

where k_2 is another unknown constant.

Again, k_2 is not specified by the model, and must (at present, at least) be discovered by trial and error (values in the range 0–2 appear about right for the TREC databases, although performance is not sensitive to this correction¹)—see the results section 7.

5.6 First Component

The first component of equation 11 is:

$$w(\underline{\mathbf{x}}, d)_1 = \log \frac{P(\underline{\mathbf{x}}|R, d) P(\underline{\mathbf{Q}}|\bar{R}, d)}{P(\underline{\mathbf{x}}|\bar{R}, d) P(\underline{\mathbf{Q}}|R, d)}.$$

¹Values of this constant depend on the base of the logarithms used in the term-weighting functions

Expanding this on the basis of term independence assumptions, and also making the assumption that eliteness is independent of document length (on the basis of the Verbosity hypothesis), we can obtain a formula for the weight of a term t which occurs tf times, as follows:

$$\begin{aligned} w(t, d)_1 &= \log \frac{(p'(\lambda d/\Delta)^{tf} e^{-\lambda d/\Delta} + (1-p')(\mu d/\Delta)^{tf} e^{-\mu d/\Delta}) (q' e^{-\lambda d/\Delta} + (1-q') e^{-\mu d/\Delta})}{(q'(\lambda d/\Delta)^{tf} e^{-\lambda d/\Delta} + (1-q')(\mu d/\Delta)^{tf} e^{-\mu d/\Delta}) (p' e^{-\lambda d/\Delta} + (1-p') e^{-\mu d/\Delta})} \\ &= \log \frac{(p' \lambda^{tf} e^{-\lambda d/\Delta} + (1-p') \mu^{tf} e^{-\mu d/\Delta}) (q' e^{-\lambda d/\Delta} + (1-q') e^{-\mu d/\Delta})}{(q' \lambda^{tf} e^{-\lambda d/\Delta} + (1-q') \mu^{tf} e^{-\mu d/\Delta}) (p' e^{-\lambda d/\Delta} + (1-p') e^{-\mu d/\Delta})}. \end{aligned} \quad (14)$$

Analysis of the behaviour of this function with varying tf and d is a little complex. The simple function used for the experiments (formula 10) exhibits some of the correct properties, but not all. In particular, 14 shows that increasing d exaggerates the S-shape mentioned in section 4.2; formula 10 does not have this property. It seems that there may be further scope for development of a rough model based on the behaviour of formula 14.

5.7 The Scope Hypothesis

As indicated above (section 5.1), an alternative hypothesis concerning document length would regard a long document as a set of unrelated, concatenated short documents. The obvious response to this hypothesis is to attempt to find appropriate boundaries in the documents, and to treat short passages rather than full documents as the retrievable units. There have been a number of experiments on these lines reported in the literature [10].

This approach appears difficult to combine with the ideas discussed above, in a way which would accommodate an explanation of document length in terms of a mixture of the two hypotheses. One possible solution is that used by Salton [11], of allowing passages to compete with full documents for retrieval. But there seems to be room for more theoretical analysis.

5.8 Document Length and Term Frequency—Summary

We have, then, a term weighting function which includes a within-document term frequency component, without (equation 8) or with (equation 10) a document-length component. We also have a document-length correction factor (equation 13) which can be applied to either 8 or 10.

6 Query Term Frequency

The natural symmetry of the retrieval situation as between documents and queries suggests that we could treat within-query term frequency (qtf) in a similar fashion to within-document term frequency. This would suggest, by analogy with equation 8, a weighting function thus:

$$w = \frac{qtf}{(k_3 + qtf)} w^{(1)}, \quad (15)$$

where k_3 is another unknown constant.

In this case, experiments (section 7) suggest a large value of k_3 to be effective—indeed the limiting case, which is equivalent to

$$w = qtf \times w^{(1)}, \quad (16)$$

appears to be the most effective. This may perhaps suggest that an S-shaped function like equation 9 could be better still, though again none has been tried in the present experiments.

The experiments are based on combining one of these qtf multipliers with the within-document term frequency and document length functions defined above. However, it should be pointed out that (a) the “natural symmetry” as between documents and queries, to which we appealed above, is open to question, and (b) that even if we accept each model separately, it is not at all obvious that they can be combined (a properly constructed combined model would have fairly complex relations between query and document terms, query and document eliteness, and relevance). Both these matters are discussed further by Robertson [12]. In the meantime, the combination of either qtf multiplier with the earlier functions must be regarded as not having a strong theoretical motivation.

7 Experiments

7.1 TREC

The TREC (Text REtrieval Conference) conferences, of which there have been two, with the third due to start early 1994, are concerned with controlled comparisons of different methods of retrieving documents from large collections of assorted textual material. They are funded by the US Advanced Projects Research Agency (ARPA) and organised by Donna Harman of NIST (National Institute for Standards and Technology). There were about 31 participants, academic and commercial, in the TREC-2 conference which took place at Gaithersburg, MD in September 1993 [2]. Information needs are presented in the form of highly structured “topics” from which queries are to be derived automatically and/or manually by participants. Documents include newspaper articles, entries from the Federal Register, patents and technical abstracts, varying in length from a line or two to several hundred thousand words.

A large number of relevance judgments have been made at NIST by a panel of experts assessing the top-ranked documents retrieved by some of the participants in TREC-1 and TREC-2. The number of known relevant documents for the 150 topics varies between 1 and more than 1000, with a mean of 281.

7.2 Experiments Conducted

Some of the experiments reported here were also reported at TREC-2 [1].

Database and Queries

The experiments reported here involved searches of one of the TREC collections, described as disks 1 & 2 (TREC raw data has been distributed on three CD-ROMs). It contains about 743,000 documents. It was indexed by keyword stems, using a modified Porter stemming procedure [13], spelling normalisation designed to conflate British and American spellings, a moderate stoplist of about 250 words and a small cross-reference table and “go” list. Topics 101–150 of the 150 TREC-1 and -2 topic statements were used. The mean length (number of unstopped tokens) of the queries derived from title and concepts fields only was 30.3; for those using additionally the narrative and description fields the mean length was 81.

Search Procedure

Searches were carried out automatically by means of City University’s Okapi text retrieval software. The weighting functions described in Sections 4–6 were implemented as BM15² (the model using equation 8 for the document term frequency component) and BM11 (using equation 10). Both functions incorporated the document length correction factor of equation 13. These were compared with BM1 ($w^{(1)}$ weights, approximately ICF, since no relevance information was used in these experiments) and with a simple coordination-level model BM0 in which terms are given equal weights. Note that BM11 and BM15 both reduce to BM1 when k_1 and k_2 are zero. The within-query term frequency component (equation 15) could be used with any of these functions.

To summarize, the following functions were used:

$$\text{(BM0)} \quad w = 1$$

$$\text{(BM1)} \quad w = \log \frac{N - n + 0.5}{n + 0.5} \times \frac{qtf}{(k_3 + qtf)}$$

$$\text{(BM15)} \quad w = \frac{tf}{(k_1 + tf)} \times \log \frac{N - n + 0.5}{n + 0.5} \times \frac{qtf}{(k_3 + qtf)} + k_2 \times nq \frac{(\Delta - d)}{(\Delta + d)}$$

$$\text{(BM11)} \quad w = \frac{tf}{\left(\frac{k_1 \times d}{\Delta} + tf\right)} \times \log \frac{N - n + 0.5}{n + 0.5} \times \frac{qtf}{(k_3 + qtf)} + k_2 \times nq \frac{(\Delta - d)}{(\Delta + d)}.$$

In the experiments reported below where k_3 is given as ∞ , the factor $qtf/(k_3 + qtf)$ is implemented as qtf on its own (equation 16).

²BM = Best Match

Evaluation

In all cases the 1000 top-ranking documents for each topic were run against the supplied relevance assessments using a standard evaluation program from the SMART projects at Cornell University. (This was the official evaluation method used in TREC-2.) The evaluation program outputs a number of standard performance measures for each query, and finally a set of measures averaged over all the queries in the run. The measures used in the tables below are average precision (AveP), precision at 5, 30 and 100 documents (P5 etc.), R-precision (RP) (precision after the number of known relevant documents for a query have been retrieved) and recall (Rcl) (final recall after 1000 documents have been retrieved).

In TREC, a distinction is made between ad-hoc (retrospective) experiments and routing (SDI or filtering). All the results reported here have been obtained using the topics and documents (and methods) used for ad-hoc experiments in TREC-2.

7.3 Results

In these experiments the weighting functions involve three parameters (k_1 , k_2 and k_3 above). It was to be expected that their values would interact, so there was a large number of preliminary runs to try to determine the extent of interaction, and to give clues as to the most fruitful ranges of values for further investigation. It should also be noted that k_1 and k_3 are parameters in expressions which are coefficients of a logarithmic value, whereas the document length correction involving k_2 is not a logarithm; hence the value of k_2 depends on the base of the logarithms used.³

Table 1 compares two of the new models with baseline $w^{(1)}$ weighting (BM1 with $k_3 = 0$), and with coordination level (BM0). For the shorter queries (title + concepts) BM15 appears somewhat better than BM1 (the difference is greater when a document length correction is added—Table 3). But its performance on the long queries is very poor. BM11 gives a very marked improvement over the baseline, particularly for the long queries.

Table 1. Comparison of term weighting functions

Function	k_1	k_2	k_3	AveP	P5	P30	P100	RP	Rcl
Query term source: titles + concepts									
BM11	1.0	0.0	0.0	0.300	0.624	0.536	0.440	0.349	0.683
BM15	1.0	0.0	0.0	0.214	0.504	0.435	0.347	0.278	0.568
BM1	0.0	0.0	0.0	0.199	0.468	0.416	0.326	0.261	0.542
BM0	—	—	—	0.142	0.412	0.336	0.270	0.209	0.411
Query term source: titles + concepts + narrative + description									
BM11	1.0	0.0	0.0	0.263	0.612	0.485	0.394	0.306	0.605
BM15	1.0	0.0	0.0	0.074	0.284	0.216	0.154	0.110	0.258
BM1	0.0	0.0	0.0	0.085	0.312	0.235	0.179	0.127	0.297
BM0	—	—	—	0.035	0.220	0.139	0.099	0.066	0.153
Database : TREC disks 1 and 2. Topics: 101-150									

Document Term Frequency

In Table 2 k_1 is varied for BM11 with the other parameters held at zero. As the value of k_1 increases from zero to about 2 performance improves sharply, then deteriorates gradually. Even very small values of k_1 have a marked effect, because the wide variation in document length implies that the term $\frac{k_1 \times d}{\Delta}$ in equation 10 can be quite large even when k_1 is as low as 0.05.

Document Length Correction

Values greater than 1 had a small detrimental effect in the BM11 model, but small positive values improved the performance of BM15, in which the document term frequency component is not normalised with respect to document length (Table 3).

³To obtain weights within a range suitable for storage as 16-bit integers, the Okapi system uses logarithms to base 2^{0.1}

Table 2. Effect of varying the document term frequency parameter k_1

Function	k_1	k_2	k_3	AveP	P5	P30	P100	RP	Rcl
Query term source: titles + concepts									
BM11	0.0	0.0	0.0	0.199	0.468	0.416	0.326	0.261	0.542
BM11	0.2	0.0	0.0	0.263	0.616	0.499	0.404	0.311	0.633
BM11	0.5	0.0	0.0	0.287	0.628	0.525	0.432	0.338	0.664
BM11	1.0	0.0	0.0	0.300	0.624	0.536	0.440	0.349	0.683
BM11	2.0	0.0	0.0	0.303	0.640	0.524	0.443	0.359	0.689
BM11	5.0	0.0	0.0	0.285	0.612	0.506	0.431	0.342	0.666
BM11	10.0	0.0	0.0	0.259	0.540	0.485	0.404	0.324	0.635
BM11	50.0	0.0	0.0	0.179	0.412	0.400	0.317	0.254	0.537
Database : TREC disks 1 and 2. Topics: 101–150									

Table 3. Effect of varying the document length correction parameter k_2

Function	k_1	k_2	k_3	AveP	P5	P30	P100	RP	Rcl
Query term source: titles + concepts									
BM11	1.0	0.0	0.0	0.300	0.624	0.536	0.440	0.349	0.683
BM11	1.0	1.0	0.0	0.300	0.632	0.532	0.438	0.349	0.678
BM11	1.0	2.0	0.0	0.296	0.632	0.523	0.437	0.347	0.671
BM11	1.0	5.0	0.0	0.275	0.624	0.506	0.424	0.335	0.631
BM15	1.0	0.0	0.0	0.214	0.504	0.435	0.347	0.278	0.568
BM15	1.0	1.0	0.0	0.229	0.512	0.453	0.364	0.292	0.598
Database : TREC disks 1 and 2. Topics: 101–150									

Query Term Frequency

Table 4 illustrates the effect of increasing k_3 from zero to infinity. For all functions and for both long and short queries performance improves fairly smoothly with k_3 . For BM11 the increase in average precision is 37% on the long queries and 12% on the short. It is to be expected that k_3 would have a greater effect with the long queries, in which there are many “noisy” terms and in which most of the more apposite terms are likely to occur more than once. However, high values of k_3 increase the performance of BM15 by 24% even on the short queries.

Table 4. Effect of varying the query term frequency parameter k_3

Function	k_1	k_2	k_3	AveP	P5	P30	P100	RP	Rcl
Query term source: titles + concepts + narrative + description									
BM11	1.0	0.0	0.0	0.263	0.612	0.485	0.394	0.306	0.605
BM11	1.0	0.0	1.0	0.299	0.648	0.510	0.428	0.341	0.659
BM11	1.0	0.0	3.0	0.331	0.640	0.540	0.456	0.378	0.712
BM11	1.0	0.0	8.0	0.346	0.656	0.549	0.466	0.390	0.737
BM11	1.0	0.0	20.0	0.351	0.644	0.560	0.471	0.395	0.744
BM11	1.0	0.0	100.0	0.353	0.648	0.565	0.473	0.396	0.745
BM11	1.0	0.0	∞	0.360	0.652	0.569	0.479	0.401	0.754
Query term source: titles + concepts									
BM11	1.0	0.0	0.0	0.300	0.624	0.536	0.440	0.349	0.683
BM11	1.0	0.0	∞	0.335	0.636	0.560	0.468	0.375	0.723
BM15	1.0	1.0	0.0	0.229	0.512	0.453	0.364	0.292	0.598
BM15	1.0	1.0	∞	0.284	0.560	0.485	0.416	0.336	0.685
BM1	0.0	0.0	0.0	0.199	0.468	0.416	0.326	0.261	0.542
BM1	0.0	0.0	∞	0.232	0.504	0.435	0.361	0.289	0.601
Database : TREC disks 1 and 2. Topics: 101–150									

7.4 Discussion

On the short queries, and without a query term frequency component, the best version of BM11 gives an increase of about 50% in average precision, with somewhat smaller improvements in the other statistics, over the baseline Robertson/Sparck Jones weighting BM1. On the long queries the proportionate improvement is very much greater still. To put this in perspective, the best results reported here (Table 4, row 8), are similar to the best reported by any of the TREC-2 participants at the time of the conference in September 1993.

Many experimental runs were also carried out on two other sets of 50 topics and on two other databases: TREC disk 3 and a subdatabase of disks 1 and 2 consisting entirely of Wall Street Journal articles. The absolute values of the statistics varied quite widely, but the rank order of treatments was very similar to those shown in the tables here.

Applicability to Other Types of Database

If documents are all the same length, BM15 is the same as BM11. The fact that BM11 was so much better than BM15 must reflect the very large variation in document length in the test collection. BM15 without a document length correction is not very much better than BM1 (Table 1, rows 2 and 3). For this reason, experiments are currently under way searching relatively short, uniform documents. Clearly, statistical characteristics of different databases (e.g. abstracts, controlled and/or free indexing) will vary widely in ways which may substantially affect the performance of the weighting functions discussed here.

Relevance Feedback

No experiments have yet been conducted with any form of relevance feedback. However, all the functions used contain the Robertson/Sparck Jones weight $w^{(1)}$, which includes relevance information if it is available. Therefore in principle all the functions could be used in a relevance feedback situation (with or without query expansion). Clearly there may be interaction between the values of the k s and the use of relevance information, which would have to be investigated.

Effect of the Verbose Topic Statements

The best results reported here (e.g. row 8 of Table 4) rely on the use of BM11 with a query term frequency component from the “long” queries. Without a qtf component results from the long queries are not good (Table 1, bottom half). Even the “short” TREC-derived queries are very long compared with the searches entered by end-users of interactive retrieval systems. However, they are not long in comparison with user need statements elicited by human intermediaries. This suggests that it might be possible to make good use, even in an interactive system, of quite long and even rambling query statements, if users can be persuaded to enter them. We might even suggest a voice-recognition system!

Individual Queries

The statistics hide the variation over individual topics. In general, different topics did best with different functions and parameters. In many cases, though, varying a parameter brings about an improvement or deterioration which is almost uniform over the set of topics. Surprisingly, we found no very large difference between the variances of the statistics obtained from different treatments.

8 Conclusions

The approach of this paper has been that a theoretical analysis of the possible effects of a variable, within the framework of the probabilistic theory of information retrieval, can give us useful insights, which can then be embodied in very much simpler formulae. This approach has shown very considerable benefits, enabling the development of effective weighting functions based on the three variables considered (term frequency within documents and queries and document length). These functions are simple extensions of the Robertson/Sparck Jones relevance weight, and therefore fit well with some other work on the development of probabilistic models.

The approach complements a number of other current approaches. In particular, it fits somewhere in between formal modelling and pure empiricism, alongside regression-based methods, to which it seems to offer some ideas.

Acknowledgements

We are very grateful to Michael Keen, Karen Sparck Jones and Peter Willett for acting as advisors to the Okapi at TREC projects. Part of this work was undertaken with the help of grants from the British Library and ARPA.

References

1. Robertson S.E. *et al.* Okapi at TREC-2. In: [2].
2. Harman D.K. (Ed.) *The Second Text REtrieval Conference (TREC-2)*. NIST Gaithersburg MD, to appear.
3. Cooper W.S. *et al.* Probabilistic retrieval in the TIPSTER collection: an application of staged logistic regression. In: Harman D.K. (Ed.) *The First Text REtrieval Conference (TREC-1)*. NIST Gaithersburg MD, 1993. (pp 73-88).
4. Croft W. and Harper D. Using probabilistic models of information retrieval without relevance information. *Journal of Documentation* 1979; 35:285-295.
5. Harter S.P. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science* 1975; 26:197-206 and 280-289.
6. Robertson S.E., Van Rijsbergen C.J. & Porter M.F. Probabilistic models of indexing and searching. In Oddy R.N. *et al.* (Eds.) *Information Retrieval Research* (pp.35-56). Butterworths London, 1981.
7. Cooper W.S. Inconsistencies and misnomers in probabilistic IR. In: *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.57-62). Chicago, 1991.
8. Robertson S.E. and Sparck Jones K. Relevance weighting of search terms. *Journal of the American Society for Information Science* 1976; 27:129-146.
9. Margulis E.L. Modelling documents with multiple Poisson distributions. *Information Processing and Management* 1993; 29:215-227.
10. Moffat A., Sacks-Davis R., Wilkinson R. & Zobel J. Retrieval of partial documents. In: Harman D.K. (Ed.) *The First Text REtrieval Conference (TREC-1)*. NIST Gaithersburg MD, 1993. (pp 59-72).
11. Buckley C., Salton G. & Allan J. Automatic retrieval with locality information using SMART. In: [2].
12. Robertson S.E. Query-document symmetry and dual models. (Unpublished.)
13. Porter M.F. An algorithm for suffix stripping. *Program* 1980; 14:130-137.