TERM FREQUENCY AND TERM VALUE

S.E. Robertson
City University, London, U.K.

It has long been recognized that the collection frequency of an index term, that is the number of documents to which it is assigned, is potentially of value in retrieval. In particular, some use has been made of term weighting functions which use (either alone or in combination with other information) term frequency data. There is much experimental evidence that such weighting schemes improve performance.

At the conference organized by this group jointly with its British counterpart, in Britain last year, Salton presented a paper (Salton and Wu, forthcoming) in which he surveyed a number of different models of the relationship between term weight and term value. He gathered together an impressive array of models to suggest that this relationship is a peaked one: that is, the best index terms are those in a middling range of frequencies; terms with frequencies outside this range are likely to be of lesser value.

Sparck Jones, in the U.K., has long been using a term-frequency based weighting function which assumes that the most valuable terms are those of least frequency. Salton presented this function as simply a reasonable approximation to the peaked curve over the most important part of the range. There is, however, some evidence for the Sparck Jones monotonic function. The object of this paper is to review this evidence and at the same time to compare the two models.

I had hoped, when I set out on this paper, to come to some firm answers. As it turns out, as we shall see, I came up only with a curious puzzle.

Some early arguments

Sparck Jones' (1972) monotonic function was initially suggested by the Zipfian charcter of the distribution of term frequencies. In particular, the distribution suggests a logarithmic function for the weighting formula. Such a function was tried by Sparck Jones and found to give improvements in retrieval performance over simple (unweighted) retrieval. The Zipf argument was not intended as a theoretical justification for the function; the only justification suggested was retrieval performance.

Salton (1975) proposed that an empirical argument should be used to determine the relationship. He measured for each term its "discrimination value", that is a measure of its ability to discriminate between relevant and non-relevant documents. Correlating this with term frequency, he observed the peaked relationship described above.

Thus far, the suggestions do not appear to be incompatible. It is possible that a very low frequency term could have low discrimination value simply because it does not occur in enough documents, but nevertheless could be a good indicator of relevance for the rare documents in which it occurs.

Probabilistic models

A new thread has been introduced into the argument recently, with the development of probabilistic models of retrieval. There are several different probabilistic models, addressing different aspects of the retrieval process, but the one which seems most appropriate to this question is that proposed by Robertson and Sparck Jones (1976).

In this model the "value" of a query term is the weight which should be assigned to the term in a simple weighted retrieval system. A probabilistic model yields (under independence assumptions) a formula for this weight, in terms of a certain probabilities, as follows:

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

where w is the weight to be assigned to a term t ("relevance weight")

p is the probability that a document d will have t as index term, given that d is relevant to the query

q is the probability that a document d will have t as index term, given that d is non relevant.

The obvious use of this model is in relevance feedback, where p and q can be estimated directly from relevance data provided by the searcher. However, there may also be indirect clues which might provide estimates of p and q prior to relevance feedback. One of these clues may be term frequency.

The Croft-Harper model

Croft and Harper (1979) derive an explicit relationship between the relevance weight of equation (1) and term frequency as follows. If we have no relevance information then there is not very much we can say about p - so it would seem appropriate to assume that p is constant (at least until such information is forthcoming). On the other hand, we can say a lot about q. Because the number of relevant documents in a collection is small compared to the total collection size (i.e. almost all documents are non-relevant), it seems likely that the relative frequency of occurrence of the term in the entire collection is a reasonable estimate of its relative frequency in the set of non-relevant documents. Thus we would get, from (1),

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

where N is the number of documents in the collection

n is the number of documents indexed by t

and p (and hence also $\log \frac{p}{1-p}$) is an unknown constant

If $p = \frac{1}{2}$, $\log \frac{p}{1-p} = 0$, so the weight reduces to $w = \log \frac{N-n}{n}$

which implies a monotonic relationship between the frequency of the term and its weight (value). In fact this formula is very close indeed to that used by Sparck Jones. For other values of p the divergence is greater, but the relationship is still monotonic.

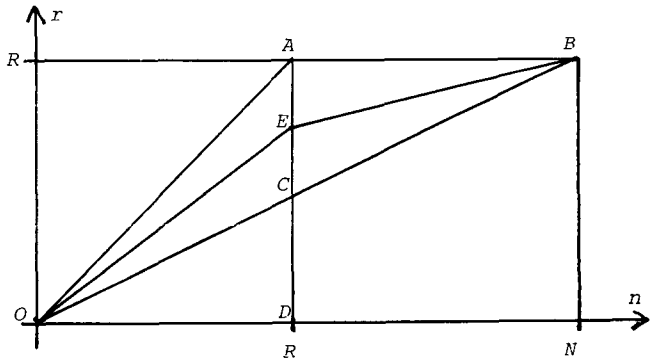
The Croft-Harper model is obviously simple and crude; in fact it is in some sense internally inconsistent. Nevertheless, it seems to provide some more direct support for the Sparck Jones formulation.

The Salton-Yu-Lam model

Selton, Yu and Lam (forthcoming) propose a somewhat more elaborate model, as follows. They postulate a relationship between the overall frequency of the term and its frequency in the set of relevant documents, according to Figure 1. (N and n are shown as above; R is the total number of relevant documents, and r is the number of relevant documents indexed by the term.

Figure 1

The Relationship Assumed by Salton, Yu and Lam



The straight line OCB represents terms of no value (retrieval using such a term is no better than retrieving a random set). The 2-segment line OAB, on the other hand, represents the best possible terms. So they assume that an average query term will lie somewhere between the two, on the two-segment line OEB.

It follows from this assumption that the value (relevance weight) of the average query term will depend on its frequency, in such a way that it reaches a peak at $n = R$ and declines to either side. Hence this model seems to support the Salton formulation.

Representing the underlying assumptions

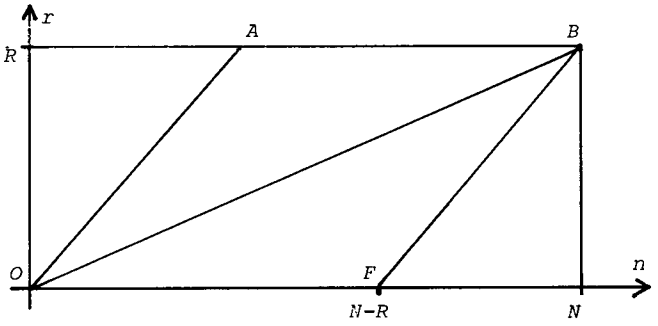
It now appears that the two formulations are strictly in conflict. Both use the same interpretation of term "value", and can claim descent from and justification in terms of the same probability model. In these circumstances, it seems desirable to represent the underlying assumptions of the alternative models in the same manner, and preferably in a manner which indicates clearly why they make different predictions.

In pursuit of this aim, it is useful to make some transformations of the graph used by Salton, Yu and Lam. These transformations are easily understood if they are broken into a few simple steps, as follows. First we mark in the worst case OFB (in addition to the best and random cases): see Figure 2(a). We now see that the full space in which we have to work is lozenge shaped (OABF). It can be made rectangular by the simple transformation of taking $n - r$ as the as the horizontal axis, instead of n: see Figure 2(b). Since $n - r$ is just the number of non-relevant documents indexed by the term, this graph is still easily interpreted.

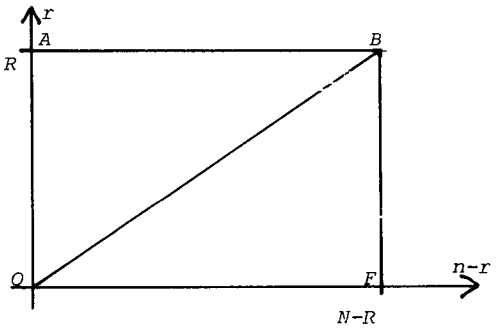
Next, we may turn the rectangle into a square, simply by normalizing the axes to r/R and $(n-r)/(N-r)$; the interpretation is still simple: see Figure 2(c). All the above transformations are lineas, which means that a straight line on the Salton-Yu-Lam graph is still a straight line on the unit square of Figure 2(c). Finally it is useful to include a non-linear transformation, taking the logistic (or log-odds) function of the proportions, giving axes log $r/(R-r)$ and log $(n-r)/(N-R-n+r)$: see Figure 2(d).

Figure 2

(a) The Salton-Yu-Lam Graph



(b) The rectangle



(c) The unit square

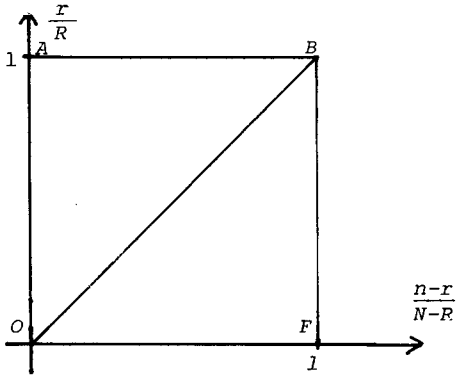
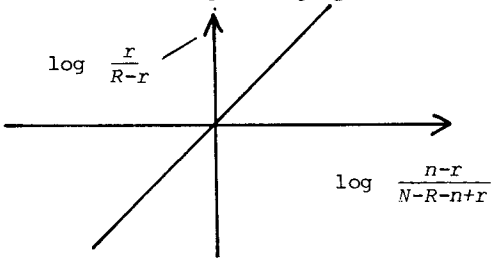


Figure 2 (continued)

(d) The logistic graph



Lines of constant value

It is appropriate to ask what shape contours representing terms of constant value would be on these various diagrams. If we ignore estimation problems (which are considered below) we have

$$p = \frac{r}{R}, \quad q = \frac{n-r}{N-R}$$

So that, from (1),

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

$$= \log \frac{r}{R-r} - \log \frac{n-r}{N-R-n+r}$$

It follows that, on the logistic graph, a curve of constant w is in fact a straight line at 45° to the axes, as in Figure 3(a). It may also be deducted that, in the unit square, the curve is as shown in Figure 3(b).

Figure 3

Constant w Contours

(a) on logistic graph

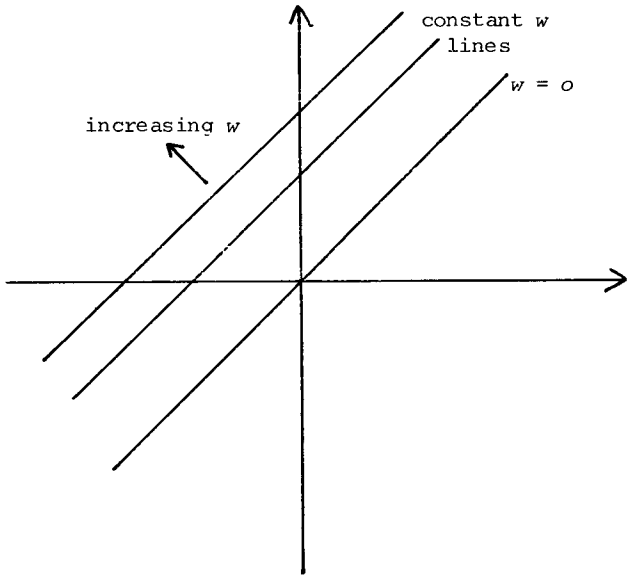


Figure 3 (continued)

(b) in unit square

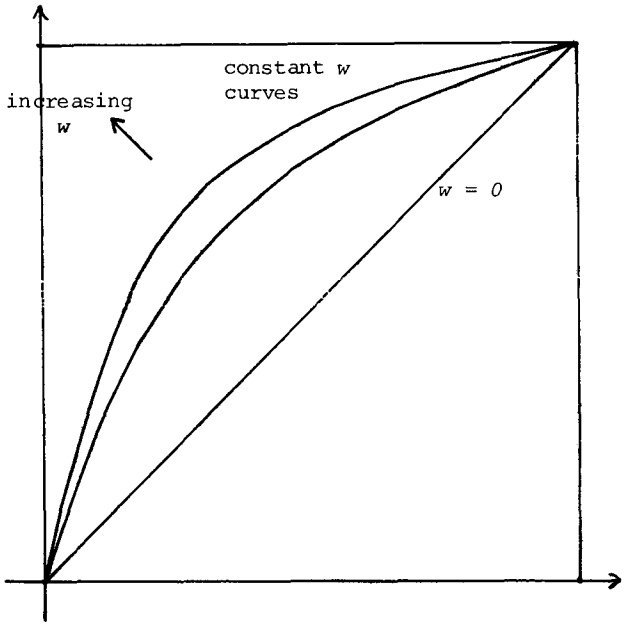
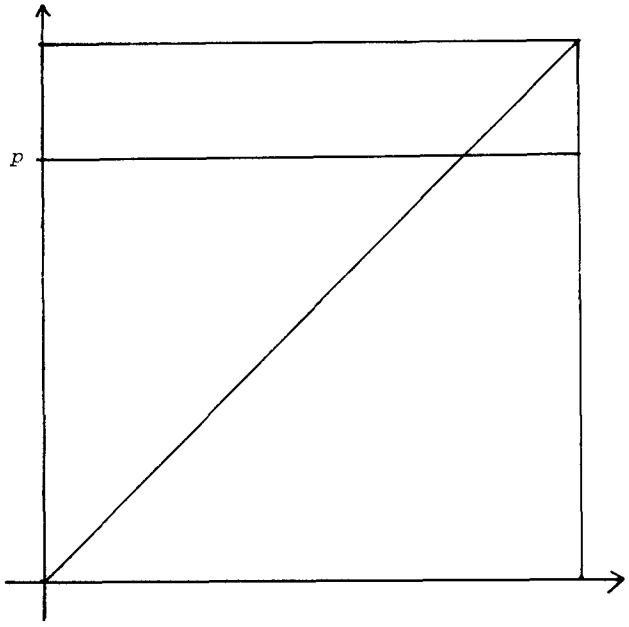


Figure 4

(a) The Croft-Harper relationship in the unit square

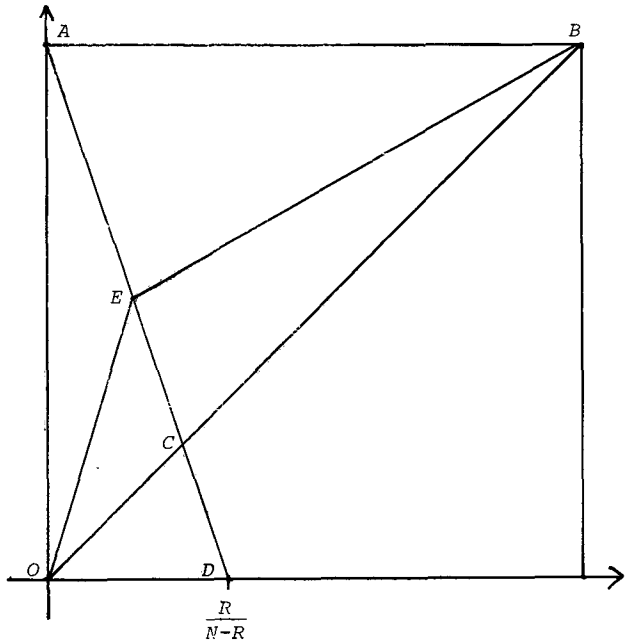


If, therefore, we assume a relationship in the form of a line on one of these graphs, then the way in which this line cuts the constant w curves will determine whether, according to our assumption, w increases or decreases with frequency. We can illustrate this point by examining the Croft-Harper and Salton-Yu-Lam models.

The Croft-Harper model assumes, simply, that p is constant - and so is represented by a horizontal line as in Figure 4(a). It is clear that w must decrease as one moves from left to right on this line (indeed it must go below zero at some point).

The Salton-Yu-Lam relationship is that given by OEB in Figure 4(b). Again, it is clear that w must increase as one moves on the straight-line segment OE, and decrease as one moves along EB.

(b) The Salton-Yu-Lam relationship in the unit square



One could clearly think of curves which predicted either relationship, according to their exact curvature at different points. Further, it could be extremely difficult to decide empirically which one gave the better fit.

This suggests that we should look for a theory - that we should seek to answer the question: "Why does such a relationship exist?" If we answer this question, we may be led to one particular form of the relationship as being more consistent with the theory than another. In order to indicate two possible explanations, I first have to discuss some earlier work (Robertson, 1976).

The linear logistic model

The following model was developed for the purpose of allowing a more sophisticated method of estimating retrieval effectiveness measures. The measures involved were recall (proportion of relevant documents retrieved) and fallout (proportion of non-relevant documents retrieved). If we substitute for "retrieved", "indexed by term t", these are the same proportions as we have identified as important in probabilistic model, namely

$$\frac{r}{R} \quad \text{and} \quad \frac{n-r}{N-R}$$

The estimation involved taking the logistic transformation of these proportions, as in Figure 2(d), and using a Bayesian estimation technique for points in this logistic space, with a bivariate normal prior distribution. The effect of this Bayesian technique was to pull the points together as far as was consistent with the data.

In fact, in almost all cases, the points were pulled onto a straight line - at less than 45° to the horizontal. (This work has some relation with the earlier work of Swets (1969) - though the transformation is somewhat different and the estimation method is very much more sophisticated.) When this line is interpreted in terms of term value, it implies a monotonic relationship between value and frequency in the manner discussed above.

The fact that the line is straight rather than curved may be an artifact of the estimation model, though some evidence discussed previously (Robertson, 1976) suggests that (a) the results for single terms are exactly parallel to those for evaluation of searches, and (b) the straight lines obtained from the evaluation of searches are not an artifact of the estimation model, but a genuine property of the results. What concerns me most here, however, is the possibility of explaining the fact that the lines are not at 45°.

Possible reasons

Is there, perhaps, some simple statistical reason for the shape of the lines? Is it in some sense to be expected on the basis of some statistical null hypothesis - a simple model involving some form of independence or similar simplifying device? The answer appears to be no - at least no such simple model exists. When Swets proposed his method for analyzing search evaluations, he first suggested (as the simplest possible model) a 45° straight line, but then was led by his results to consider a more complex model (Swets, 1969). This more complex model is purely descriptive (has no explanatory power), and besides has some theoretical deficiencies (Robertson, 1977); the sole point of introducing it was to deal with the observation of the non-45° lines.

I also started with a simple 45° model, and had to abandon it in the face of evidence. So I then looked for further explanations.

Retrievability

One possibility is that the relationship has to do with the variations in retrievability in the relevant document set. We can set up a model which is as simple as the 45° model statistically, but which allows for the fact that some relevant documents are inherently more likely to contain terms by which they can be retrieved than others. This property has been observed empirically before; it can be related to degree of relevance, though there may be other reasons for it.

To take a very simple case, suppose there are just two classes of relevant documents, those that are easy to retrieve and those that are hard to retrieve, but that each set has its own 45° line. Then it can be shown (Robertson, 1976) that the composite set has a curve which approximates (over the middle of the range) to a less-than-45° line. Thus this property of relevant documents may be an explanation for the relationship. I originally concluded that it did not account for the full extent of the relationship, but a number of my assumptions might be questioned, and it may be possible to provide a full explanation on these lines. On the other hand, a similar variation among the non-relevant documents would produce the opposite effect.

Estimation problems

A further possibility is that any empirically observed relationship is an artifact of the method of estimating the relevance weight. It might be argued that the empirical investigation of this area does not require any estimation - that if we have complete relevance information, we can assign "true" values to p and q (and thus to w). This suggestion runs up against theoretical and practical difficulties, both of which are associated with extreme values of p and q. Suppose, for example, the term in question occurs in all the relevant documents (so that $p = 1$ is the "true" value). In what sense can we regard this value as suitable for the sort of analysis we have in mind?

The practical difficulty is simply stated - if we set $p = 1$, we get value of infinity for w, which is unfortunately somewhat difficult to manipulate. The theoretical difficulty lies in

the associated idea that if $p = 1$, the term is in some sense perfect. It seems to me that we cannot reasonably regard this as a genuine property of the term - it must be an accident of the particular document collection. Which means that we must regard the document collection as a sample from some (actual or notional) larger population. Which in turn implies that we do not have complete information about p at all - what we have is sample information, from which the true value must be estimated.

Even if we accept that we must regard the process as one of estimation, it would still be possible to estimate p and q by simple proportions But there are still difficulties. The practical difficulty remains the same (that of how to deal with infinities); the theoretical difficulties include the same problem of interpretation, but also a new dimension, that of bias. Because the logistic function is non-linear, a good (in the sense of unbiased) estimator of a probability does not yield a good estimator of the logistic transformation of the probability.

So we are forced to consider other methods of estimation. In this case, it appears, all other methods are explicitly or implicitly Bayesian. (Bayesian methods, for those not familiar with the term, assume some prior expectation of the parameter to be estimated and modify this prior expectation gradually as more data is acquired.) The problem is that it is very difficult to investigate the relationship empirically without artificially introducing it into the situation by way of the prior expectations.

This problem emerged very clearly with the explicitly Bayesian estimation method used in the earlier study; after some simulation experiments, I concluded that it almost certainly was to blame for some of the deviation from 45°, but not for all of it. (Again, however, some of the assumptions used in the simulation are questionable.) What is perhaps not so obvious is that the same problem comes into play with other, simpler estimation methods.

Consider, for example, the estimation formula originally proposed for the Robertson-Sparck Jones weighting function (1), and since used by (among others) Salton, Yu and Lam. The fomula is:

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

This is known as the "point-five fomula" (compare (2), which is the same without the 0 5's). These 0.5's were added in order to minimize bias (Robertson and Sparck Jones, 1976), but they also serve a Bayesian function (van Rijsbergen, 1977). This can be illustrated with two extreme cases.

Suppose $N = 1000$ and $R = 10$. If the term occurs in no documents at all it will be given a weight of

$$w = \log \frac{990.5}{10.5} \approx 2$$

If, on the other hand, it occurs in all the documents in the collection, it will be given weight

$$w = \log \frac{10.5}{990.5} \approx -2$$

although in neither case is there any empirical justification for the weight.

Salton, Yu and Lam (forthcoming), in reporting some experimental evidence to support the predictions of their model, in fact use the point-five formula. The above considerations somewhat reduce the apparent value of their results.

Experimental results

In the course of developing the probabilistic model, we have devised an estimation method for the relevance weights which is explicitly Bayesian but which does not, as the earlier one did, take into consideration the weights of other terms when estimating the weight of a particular one. (This method will be more fully reported elsewhere.) Furthermore, it is possible to provide it with different Bayesian prior distributions. For this experiment, therefore, we start with a prior which does not depend on term frequency.

The experiment was performed on the National Physical Laboratory test collection. It is worth mentioning, first, that in this collection (which is rather larger than anything used by Salton, Yu and Lam), there are very few terms indeed whose frequencies are less than the numbers of relevant documents for the queries in which they occur. Thus the peak at $n = R$ predicted by Salton, Yu and Lam would be very difficult to establish or refute. It seems likely that, in this respect, the NPL collection is more realistic than the smaller ones.

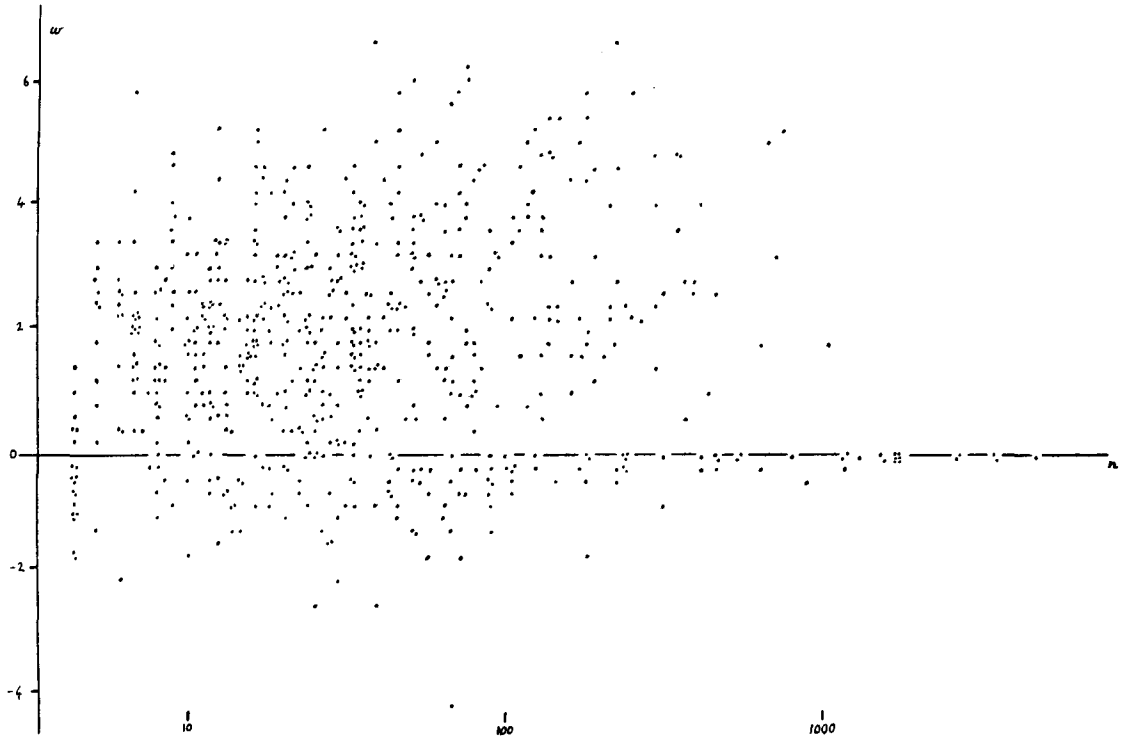
The results of the experiment are displayed as a scatter diagram in Figure 5; averages for groups of terms by frequency are shown in Figure 6.

The immediate reaction to these results must be that there is little or no relationship discernible. There could be a suggestion of a peaked relationship, though when we look more closely at the terms with near-zero weights at the high-frequency end, most of them turn out to be words which should have been on a stop-list (the indexing of the NPL collection was done automatically from abstracts). Making a slightly longer stop-list almost completely eliminates the drop at the high-frequency end.

The results suggest that earlier observations of a relationship have been completely the result of the estimation method. However, such a conclusion does not square with the retrieval test results, which show universally (with the NPL collection as with others) that reducing the weight of the high-frequency terms improves performance.

Figure 5

Value by Frequency for 650 Query Terms from the NPL Collection



So we are left with something of a puzzle. Some possible but more complex ideas suggest themselves - e.g. that the explanation for the value of frequency weighting does not lie in the Robertson-Sparck Jones probabilistic model, because (for example) it involves interactions between terms (where the Robertson-Sparck Jones model assumes independence). Clearly, further investigation is required.

Conclusions

Neither the theoretical models, nor the experimental investigations, discussed so far in the literature have provided a convincing account of the relationship between term value and term frequency. This relationship is, however, of considerable interest, both from the point of view of devising good retrieval rules, and because the explanation could tell us something quite fundamental about retrieval characteristics of documents. Further investigation of the problem is recommended.

Acknowledgements

John Bovey devised the new estimation method and performed the analysis. The test collection was made available by the National Physical Laboratory, Teddington, England.

References

W.B. Croft and D.J. Harper, Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285-279, 1979.

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

S.E. Robertson, A theoretical model of the retrieval characteristics of information retrieval systems. Ph.D. Thesis, University of London, 1976.

S.E. Robertson, Progress in documentation: Theories and models in information retrieval. Journal of Documentation, 33, 126-148, 1977.

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

G. Salton, A Theory of indexing. Regional Conference Series in Applied Mathematics. Philadelphia: SIAM, 1975.

G. Salton and H. Wu. Paper presented at the BCS-ACM Conference on Research and Development in Information Retrieval, Cambridge, June 1980. To appear in: Information Retrieval Research, Butterworths, forthcoming.

G. Salton, C.T. Yu and K. Lam, Optimum term weighting. Journal of the ACM (forthcoming).

K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 11-21, 1972.

J.A. Swets, Effectiveness of information retrieval methods. American Documentation, 20, 72-89, 1969.

Figure 6

Average Value for Terms Grouped by Frequency