Reprinted from: *Journal of Documentation* **28** no. 2, pp 164–165

Dear Sir,

TERM SPECIFICITY

May I comment on a small point in Sparck Jones' otherwise excellent article in your March issue?

Sparck Jones proposes that terms used in requests in a co-ordinate indexing system should be weighted according to the frequencies of occurrence of the terms in the collection. The function $f(n)$ used to determine the weights is defined as follows:

$$f(n) = m \quad \text{where } 2^{m-1} < n \leq 2^m$$

In fact, one can rewrite this definition as

$$f(n) \approx \log_2 n \quad \text{logarithm to the base 2 of } n,$$

where the approximation made is to take the next higher integer.

Then if there are N documents in the collection, the weight of a term which occurs n times is defined by Sparck Jones as:

$$\begin{aligned} f(N) - f(n) + 1 &\approx \log N - \log n + 1 \\ &\approx \log(N/n) + 1 \end{aligned}$$

Why the +1? It would seem more logical to use $\log(N/n)$—for a term which was used to index every item in the collection (obviously useless for retrieval), $\log(N/n)$ would give a weight of 0, whereas the formula above gives a weight of 1. I suspect the answer to this question lies in the approximation mentioned above: without the +1, the approximation would have the effect of giving a number of other terms zero weight (e.g. in the Cranfield 200 collectin, any term for which $n > 128$. In fact, however, this situation occurs only once in the three collections considered ('Flow-' in the Cranfield collection), and probably only occurs here because this collection is such a specialised subset of the original Cranfield 1400 document collection.

There are also theoretical arguments for using $\log(N/n)$, which might serve to shed some light on Sparck Jones' observations (the +1 is probably not in fact very important). The ration n/N is the proportion of items in the collection in which the term occurs - i.e. the probability (say p) that a given item (chosen at random) will contain the term. Then the weight of the term is $\log(1/p)$. Suppose that an item contains the terms a,b,c in common with the question; let the values of p for these terms be p_a, p_b, p_c respectively. Then the weight ('level') assigned to the document is

$$\log(1/p_a) + \log(1/p_b) + \log(1/p_c) = \log(1/p_a p_b p_c)$$

Now, $p_a p_b p_c$ can be interpreted as the probability that a document will randomly contain all three terms a,b,c. Therefore the use of the weight $\log(N/n)$ is a quantification of the statemenbt: 'The less likely (on a random basis) it is that a given combination of terms occurs, the more likely it is that a document containing this combination is relevant to the question.'

That this statement turns out to be a better basis for retrieval than the corresponding assumption which is the basis for the usual 'level of co-ordination' (that the probability of relevance is simple related to the number of terms in common with the question) is hardly surprising. This does not, of course, detract from the value of having the proposition demonstrated in practice.

Sincerely,
S. E. ROBERTSON
Research and Development Department[1]

Dr Sparck Jones writes: The formula given was in fact used as a convenient means of computing the logarithm mentioned by Mr Robertson. It was perhaps an oversight to give the algorithm rather than the basis for it. the reasons for using a logarithmic weighting are as Mr Robertson says: his theoretical argument is quite right and in any case a logarithmic weighting is intuitively the obvious one. As for the '+1', Mr Robertson is again right. One is reluctant to discard terms altogether unless they occur in all or nearly all of the documents. In fact, 'flow-' is a popular Cranfield request term, and it would be a mistake to reject it altogether, when it does carry some information, though admittedly not too much. On Mr Robertson's last point, what is really surprising is that this obvious notion does not seem to be widely implemented in mechanised retrieval systems.

[1] of Aslib, the then publishers of *Journal of Documentation*