A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL

KAREN SPARCK JONES
University of Cambridge Computer Laboratory

The exhaustivity of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning. The effects on retrieval of variations in term specificity are examined, experiments with three test collections showing in particular that frequently-occurring terms are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms. Results for the test collections show that considerable improvements in performance are obtained with this very simple procedure.

EXHAUSTIVITY AND SPECIFICITY

WE ARE FAMILIAR with the notions of exhaustivity and specificity: exhaustivity is a property of index descriptions, and specificity one of index terms. They are most clearly illustrated by a simple keyword or descriptor system. In this case the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it; and the specificity of an individual term is the level of detail at which a given concept is represented.

These features of a document retrieval system have been discussed by Cleverdon *et al.*[4] and Lancaster,[6] for example, and the effects of variation in either have been noted. For instance, if the exhaustivity of a document description is increased by the assignment of more terms, when the number of terms in the indexing vocabulary is constant, the chance of the document matching a request is increased. The idea of an optimum level of indexing exhaustivity for a given document collection then follows: the average number of descriptors per document should be adjusted so that, hopefully, the chances of requests matching relevant documents are maximized, while too many false drops are avoided. Exhaustivity obviously applies to requests too, and one function of a search strategy is to vary request exhaustivity. I shall be mainly concerned here, however, with document descriptions.

Specificity as characterized above is a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise. This is a natural view for anyone concerned with the construction of

an entire indexing vocabulary. Some decision has to be made about the discriminating power of individual terms in addition to their descriptive propriety. For example, the index term 'beverage' may be as properly used for documents about tea, coffee, and cocoa as the terms 'tea', 'coffee', and 'cocoa'. Whether the more general term 'beverage' only is incorporated in the vocabulary, or whether 'tea', 'coffee', and 'cocoa' are adopted, depends on judgements about the retrieval utility of distinctions between documents made by the latter but not the former. It is also predicted that the more general term would be applied to more documents than the separate terms 'tea', 'coffee', and 'cocoa', so the less specific term would have a larger collection distribution than the more specific ones.

It is of course assumed here that such choices when a vocabulary is constructed are exclusive: we may either have 'beverage' or 'tea', 'coffee', and 'cocoa'. What happens if we have all four terms is a different matter. We may then either interpret 'beverage' to mean 'other beverages' or explicitly treat it as a related broader term. I shall, however, disregard these alternatives here.

In setting up an index vocabulary the specificity of index terms is looked at from one point of view: we are concerned with the probable effects on document description, and hence retrieval, of choosing particular terms, or rather of adopting a certain set of terms. For our decisions will in part be influenced by relations between terms, and how the set of chosen terms will collectively characterize the set of documents. But throughout we assume some level of indexing exhaustivity. We are concerned with obtaining an effective vocabulary for a collection of documents of some broadly known subject matter and size, where a given level of indexing exhaustivity is believed to be sufficient to represent the content of individual documents adequately, and distinguish one document from another.

Index term specificity must, however, be looked at from another point of view. What happens when a given index vocabulary is actually used? We predict when we opt for 'beverage', for example, that it will be used more than 'cocoa'. But we do not have much idea of how many documents there will be to which 'beverage' may appropriately be assigned. This is not simply determined even when some level of exhaustivity is assumed. There will be some documents which cry out for 'beverage', so to speak, and we may have some idea of what proportion of the collection this is likely to be. There will also be documents to which 'beverage' cannot justifiably be assigned, and this proportion may also be estimated. But there is unfortunately liable to be some number of documents to which 'beverage' may or may not be assigned, in either case quite plausibly. In general, therefore, the actual use of a descriptor may diverge considerably from the predicted use. The proportions of a collection to which a term does and does not belong can only be estimated very roughly; and there may be enough intermediate documents for the way the term is assigned to these to affect its overall

distribution considerably. Over a long period the character of the collection as a whole may also change, with further effects on term distribution.

This is where the level of exhaustivity of description matters. As a collection grows maintaining a certain level of exhaustivity may mean that the descriptions of different documents are not sufficiently distinguished, while some terms are very heavily used. More generally, great variation in term distribution is likely to appear. It may thus be the case that a particular term becomes less effective as a means of retrieval, whatever its actual meaning. This is because it is not discriminating. It may be properly assigned to documents, in the sense that their content justifies the assignment; but it may no longer be sufficiently useful in itself as a device for distinguishing the typically small class of documents relevant to a request from the remainder of the collection. A frequently used term thus functions in retrieval as a non-specific term, even though its meaning may be quite specific in the ordinary sense.

STATISTICAL SPECIFICITY

It is not enough, in other words, to think of index term specificity solely in setting up an index vocabulary, as having to do with accuracy of concept representation. We should think of specificity as a function of term use. It should be interpreted as a statistical rather than semantic property of index terms. In general we may expect vaguer terms to be used more often, but the behaviour of individual terms will be unpredictable. We can thus redefine exhaustivity and specificity for simple term systems: the exhaustivity of a document description is the number of terms it contains, and the specificity of a term is the number of documents to which it pertains. The relation between the two is then clear, and we can see, for instance, that a change in the exhaustivity of descriptions will affect term specificity: if descriptions are longer, terms will be used more often. This is inevitable for a controlled vocabulary, but also applies if extracted keywords are used, particularly in stem form. The incidence of words new to the keyword vocabulary does not simply parallel the number of documents indexed, and the extraction of more keywords per document is more likely to increase the frequency of current keywords than to generate new ones.

Once this statistical interpretation of specificity, and the relation between it and exhaustivity, are recognized, it is natural to attempt a more formal approach to seeking an optimum level of specificity in a vocabulary and an optimum level of exhaustivity in indexing, for a given collection. Within the broad limits imposed by having sensible terms, i.e. ones which can be reached from requests and applied to documents, we may try to set up a vocabulary with the statistical properties which are hopefully optimal for retrieval. Purely formal calculations may suggest the correct number of terms, and of terms per document, for a certain degree of document discrimination. Work on these lines has been done by Zunde and Slamecka,[10]

for instance. More informally, the suggestion that descriptors should be designed to have approximately the same distribution, made by Salton for example, is motivated by respect for the retrieval effects of purely statistical features of term use.

Unfortunately abstract calculations do not select actual terms. Nor are document collections static. More importantly, it is difficult to control requests. One may characterize documents with a view to distinguishing them nicely and then find that users do not provide requests utilizing these distinctions. We may therefore be forced to accept a de facto non-optimal situation with terms of varying specificity and at least some disagreeably non-specific terms. There will be some terms which, whatever the original intention, retrieve a large number of documents, of which only a small proportion can be expected to be relevant to a request. Such terms are on the whole more of a nuisance than rare, over-specific terms which fail to retrieve documents.

These features of term behaviour can be illustrated by examples from three well-known test collections, obtained from the Aslib Cranfield, INSPEC, and College of Librarianship Wales projects. In fact in these the vocabulary consists of extracted keyword stems, which may be expected to show more variation than controlled terms. But there is no reason to suppose that the situation is essentially different. Full descriptions of the collections are given in Cleverdon *et al*,[4] Aitchison *et al*,[1] and Keen (forthcoming). Relevant characteristics of the collections are given in Section A of Table 1. The INSPEC Collection, for instance, has 541 documents in-

TABLE 1

	Cranfield	*INSPEC*	*Keen*
A. Number of documents	200	541	797
Number of terms	712	1,341	939
Number of terms per document	32	12·2	7·9
Number of documents per term	9	4·9	6·1
B. Number of requests	42	97	63
Number of terms represented	166	248	183
Number of terms per request	6·9	5·6	5·3
Number of documents per request term	31·6	11·5	44·8
C. Number of retrieving terms per request	5	3·2	3·3
Number of retrieving terms per document	1·8	1·2	1·2
Number of retrieving terms per relevant document	3·6	2	1·8
D. Number of frequent terms	96	73	50
Number of frequent terms per request	4	2·5	2·3

dexed by 1,341 terms. In all the collections, there are some very frequently occurring terms: for example in the Cranfield collection, one term occurs in 144 out of 200 documents; in the INSPEC one term occurs in 112 out of 541, and in the Keen collection one term occurs in 199 out of 797 documents. The terms concerned do not necessarily represent concepts central to the subject areas of the collections, and they are not always general terms. In the Keen collection, which is about information science, the most frequent term is 'index–', and other frequent ones include 'librar–', 'inform–', and 'comput–'. In the INSPEC collection the most frequent is 'theor–', followed by 'measur–' and 'method–'. And in the Cranfield collection the most frequent is 'flow–', followed by 'pressur–', 'distribut–' and 'bound–' (boundary). The rarer terms are a fine mixed bag including 'purchas–', and 'xerograph–' for Keen, 'parallel–' and 'silver–' for INSPEC, and 'logarithm–' and 'seri–' (series) for Cranfield.

SPECIFICITY AND MATCHING

How should one cope with variable term specificity, and especially with insufficiently specific terms, when these occur in requests? The untoward effects of frequent term use can in principle be dealt with very naturally, through term combinations. For instance, though the three terms 'bound–', 'layer–', and 'flow–' occur in 73, 62, and 144 documents each in the Cranfield collection, there are only fifty documents indexed by all three terms together. Relying on term conjunction is quite straightforward. It is in particular a way of overcoming the untoward consequences of the fact that requests tend to be formulated in better known, and hence generally more frequent, terms. It is unfortunate, but not surprising, that requests tend to be presented in terms with an average frequency much above that for the indexing vocabulary as a whole. This holds for all three test collections, as appears in Section B of Table 1. For the Cranfield collection, for example, the average number of postings for the terms in the vocabulary is nine, while the average for the terms used in the requests is 31·6; for Keen the figures are 6·1 and 44·8.

But relying on term combination to reduce false drops is well-known to be risky. It is true that the more terms in common between a document and a request, the more likely it is that the document is relevant to the request. Unfortunately, it just happens to be difficult to match term conjunctions. This is well exhibited by the term matching behaviour of the three collections, as shown in Section C of Table 1. The average number of starting terms per request ranges from 5·3 for Keen to 6·9 for Cranfield. But the average number of retrieving terms per request, i.e. the average of the highest matching scores, ranges from 3·2 to 5·0. More importantly, the average number of matching terms for the relevant documents retrieved ranges from only 1·8 for Keen to 3·6 for Cranfield, though fortunately

the average for all documents retrieved, which are predominantly non-relevant, ranges from a mere 1·2 to 1·8.

Clearly, one solution to this problem is to provide for more matching terms in some way. This may be achieved either by providing alternative substitutes for given terms, through a classification; or by increasing the exhaustivity of document or request specifications, say by adding statistically associated terms. But either approach involves effort, perhaps considerable effort, since the sets of terms related to individual terms must be identified. The question naturally arises as to whether better use of existing term descriptions can be made which does not involve such effort.

As very frequently occurring terms are responsible for noise in retrieval, one possible course is simply to remove them from requests. The fact that this will reduce the number of terms available for conjoined matching may be offset by the fact that fewer non-relevant documents will be retrieved. Unfortunately, while frequent terms cause noise, they are also required for reasonably high recall. For all three test collections, the deletion of very frequent terms by the application of a suitable threshold leads to a decline in overall performance. For the INSPEC collection, for example, the threshold was set to delete terms occurring in twenty or more documents, so that seventy-three terms out of the total vocabulary of 1,341 were removed. The effect in retrieval performance is illustrated by the recall/precision graph of Figure 1 for the Cranfield collection. Matching is by simple term

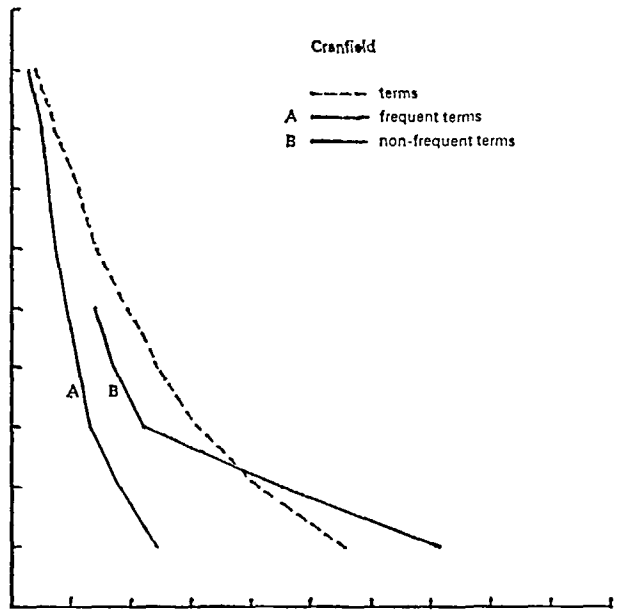

FIG. 1

co-ordination levels, and averaging over the set of requests is by straightforward average of numbers. Precision at ten standard recall values is then interpolated. The same relationship between full term matching and this restricted matching with non-frequent terms only is exhibited by the other collections: the recall ceiling is lowered by at least 30%, and indeed for the Keen collection is reduced from 75% to 25%, though precision is maintained.

Inspection of the requests shows why this result is obtained. Not merely is request term frequency much above average collection frequency; the comparatively small number of very frequent terms plays a large part in request formulation. 'Flow–' for example, appears in twelve Cranfield requests out of forty-two, and in general for all three collections about half the terms in a request are very frequent ones, as shown in Section D of Table 1. Throwing very frequent terms away is throwing the baby out with the bath water, since they are required for the retrieval of many relevant documents. The combination of non-frequent terms is discriminating, but no more than that of frequent and non-frequent terms. The value of the non-frequent terms is clearly seen, on the other hand, when matching using frequent terms only is compared with full matching, also shown in Figure 1. Matching levels for total and relevant documents are nearly as high as for all terms, but the non-frequent terms in the latter raise the relevant matching level about 1.

These features of term retrieval suggest that to improve on the initial full term performance we need to exploit the good features of very frequent and non-frequent terms, while minimizing their bad ones. We should allow some merit in frequent term matches, while allowing rather more in non-frequent ones. In any case we wish to maximize the number of matching terms.

WEIGHTING BY SPECIFICITY

This clearly suggests a weighting scheme. In normal term co-ordination matches, if a request and document have a frequent term in common, this counts for as much as a non-frequent one; so if a request and document share three common terms, the document is retrieved at the same level as another one sharing three rare terms with the request. But it seems we should treat matches on non-frequent terms as more valuable than ones on frequent terms, without disregarding the latter altogether. The natural solution is to correlate a term's matching value with its collection frequency. At this stage the division of terms into frequent and non-frequent is arbitrary and probably not optimal: the elegant and almost certainly better approach is to relate matching value more closely to relative frequency. The appropriate way of doing this is suggested by the term distribution curve for the vocabulary, which has the familiar Zipf shape. Let $f(n) = m$ such that $2^{m-1} < n \leqslant 2^m$. Then where there are N documents in the collection, the weight of a term which

occurs n times is $f(N) - f(n) + 1$. For the Cranfield collection with 200 documents, for example, this means that a term occurring ninety times has weight 2, while one occurring three times has weight 7.

The matching value of a term is thus correlated with its specificity and the retrieval level of a document is determined by the sum of the values of its matching terms. Simple co-ordination levels are replaced by a more sophisticated quasi-ranking. The effect can be illustrated by the different retrieval levels at which two documents matching a request on the same number of relatively frequent and relatively non-frequent terms respectively. With the Cranfield range of values, a document matching on two terms with frequencies 15 and 43 will be retrieved at level $5 + 3 = 8$, while one matching on terms with frequencies 3 and 7 will be retrieved at level $7 + 6 = 13$. Clearly, as the range of levels is 'stretched', more discrimination is possible.

The idea of term weighting is not new. But it is typically related to the presumed importance of a term with respect to a document in itself. For instance, if a document is mainly about paint and only mentions varnish in passing, we may utilize some simple weighting scale to assign a weight of 2 to the term 'paint' and 1 to 'varnish'. More informally, in putting a request, we may state that during searching term x must be retained, but term y may be dropped. More systematic weighting on a statistical base may be adopted if the necessary information is available. If the actual frequency of occurrence of terms in a document (or abstract) is known, this may be used to generate weights. Artandi and Wolfe[2] report the use of frequency to select a weight from a three-point scale, while Salton[7] more wholeheartedly uses the frequency of occurrence as a weight. In a range of experiments Salton has demonstrated that weighting terms in this way leads to a noticeable improvement in performance over that obtained for unweighted terms.

Weighting by collection frequency as opposed to document frequency is quite different. It places greater emphasis on the value of a term as a means of distinguishing one document from another than on its value as an indication of the content of the document itself. The relation between the two forms of weighting is not obvious. In some cases a term may be common in a document and rare in the collection, so that it would be heavily weighted in both schemes. But the reverse may also apply. It is really that the emphasis is on different properties of terms.

The treatment of term collection frequency in connection with term matching does not seem to have been systematically investigated. The effect of term frequency on statistical associations has been studied, for example by Lesk, but this is a different matter. The fact that a given term is likely to retrieve a large number of documents may be informally exploited in setting up searches, in particular in the context of on-line retrieval as described by Borko[5] for example. More whole-hearted approaches are prob-

ably hampered by the lack of the necessary information. Such a procedure as the one described is also much more suited to automatic than manual searching. It is of interest, therefore, that term frequencies have been exploited in the general manner indicated within an operational interactive retrieval system for internal reports implemented at A. D. Little (Curtice and Jones[5]). In this system indexing keywords are extracted automatically from text, and the weighting is therefore associated with a changing vocabulary and collection. However, no systematic experiments are reported.

EXPERIMENTAL RESULTS

The term weighting system described was tried on the three collections. As noted, these are very different in character, with different sizes of vocabulary, document description, and request specification, as indicated in Table 1. In all cases, however, matching with term weighting led to a substantial improvement in performance over simple term matching. The results presented in the form mentioned earlier, are given in Figure 2. A simple significance test based on the difference in area enclosed by the curves shows that the improvement given by weighted terms is fully significant, the difference being well above the required minimum.

These results are of interest for two reasons. All three collections have been used for a whole range of experiments with different index languages,

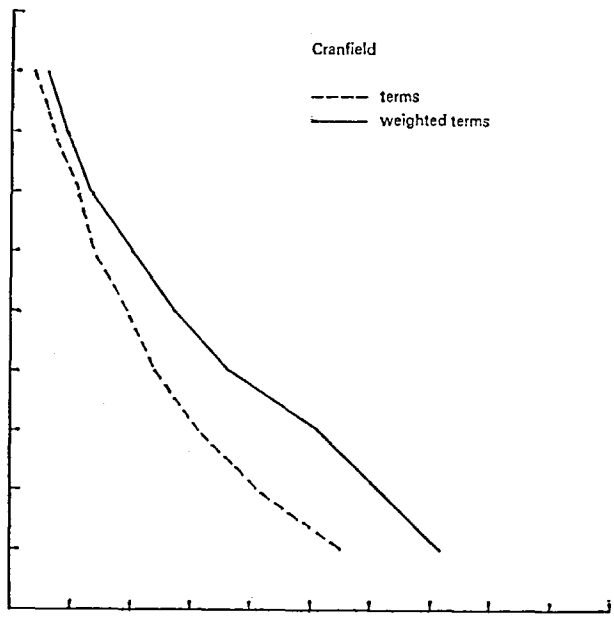

FIG. 2*a*

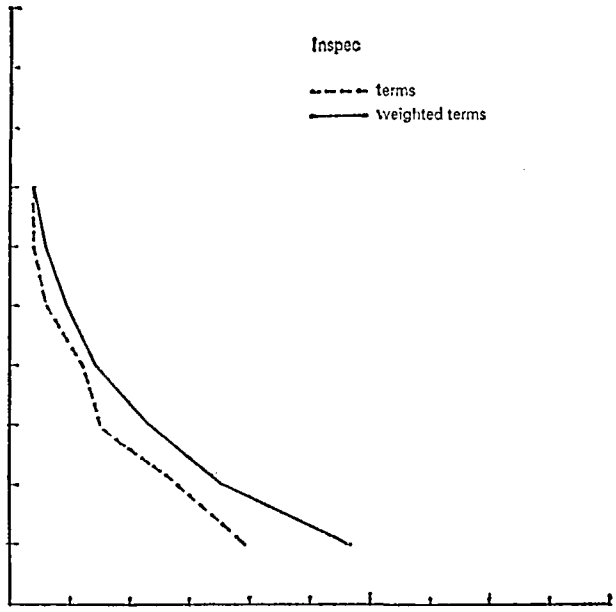

FIG. 2*b*

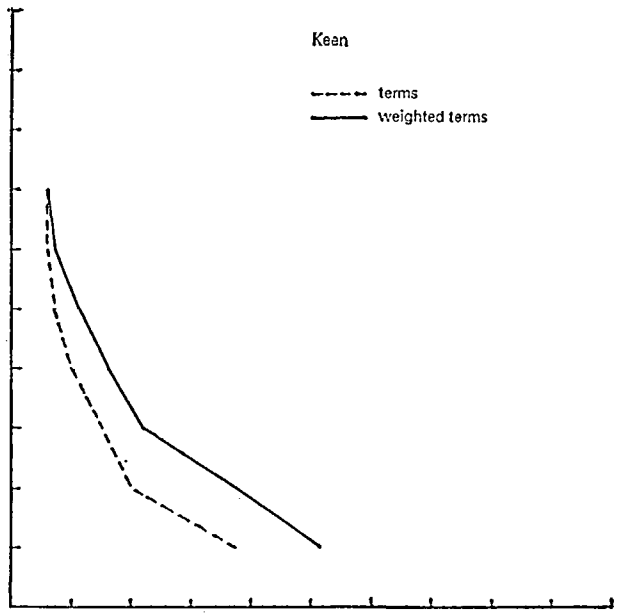

FIG. 2*c*

search techniques, and so on: see Cleverdon *et al*,[4] Salton,[7] Salton and Lesk,[8] and Sparck Jones;[9] Aitchison *et al*[1] and Keen (forthcoming). The performance improvement obtained here nevertheless represents as good an improvement over simple unweighted keyword stem matching as has been obtained by any other means, including carefully constructed thesauri: Salton's iterative search methods are not comparable. The details of the way these experimental results are presented varies, so rigorous comparisons are impossible: but the general picture is clear. Indeed, in so far as anything can be called a solid result in information retrieval research, this is one. The second point about the present results is that the improvement in performance is obtained by extremely simple means. It is compatible with an initially plain method of indexing, namely the use of extracted keywords, which may be reduced to stems automatically; it is readily implemented given an automatic term-matching procedure, since all that is required is a term frequency list and this is easily obtained; and it has the merit that the weight assigned to terms is naturally adjusted to follow the growth of and changes in a collection. Experiments with very much larger collections than those used here are clearly desirable; they will hopefully not be long delayed.

REFERENCES

1. AITCHISON, T. M., HALL, A. M., LAVELLE, K. H., *and* TRACY, J. M. *Comparative evaluation of index languages, Part II; Results*, Project INSPEC, Institute of Electrical Engineers, London, 1970.
2. ARTANDI, S. *and* WOLF, E. H. The effectiveness of automatically-generated weights and links, *American Documentation*, vol. 20, 1969, p. 198–202.
3. BORKO, H. Interactive document storage and retrieval systems—design concepts, *Mechanised information storage, retrieval and dissemination* (ed. Samuleson), Amsterdam, North-Holland, 1968, p. 591–99.
4. CLEVERDON, C. W., MILLS, J., *and* KEEN, E. M. *Factors determining the performance of indexing systems*, 2 vols. Cranfield, 1966.
5. CURTICE, R. M. *and* JONES, P. E. An operational interactive retrieval system. Cambridge, Mass., Arthur D. Little Inc., 1969.
6. LANCASTER, F. W., *Information retrieval systems: characteristics, testing and evaluation*, New York, Wiley, 1968.
7. SALTON, G. *Automatic information organization and retrieval*, New York, McGraw-Hill, 1968.
8. SALTON, G. *and* LESK, M. E. Computer evaluation of indexing and text processing, *Journal of the ACM*, vol. 15, 1968, p. 8–36.
9. SPARCK JONES, K. *Automatic keyword classification for information retrieval*, London, Butterworths, 1971.
10. ZUNDE, P. *and* SLAMECKA, V. Distribution of indexing terms for maximum efficiency of information transmission, *American Documentation*, vol. 18, 1967, p. 104–8.