DOCUMENTATION NOTE

ON TERM SELECTION FOR QUERY EXPANSION

S. E. ROBERTSON

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

In the framework of a relevance feedback system, term values or term weights may be used to (a) select new terms for inclusion in a query, and/or (b) weight the terms for retrieval purposes once selected. It has sometimes been assumed that the same weighting formula should be used for both purposes. This paper sketches a quantitative argument which suggests that the two purposes require different weighting formulae.

1. INTRODUCTION

Term weighting

Various formulae have been proposed or used, at various times, to quantify the value or usefulness of a search term in retrieval. The motivation or justification for using a particular formula may be based in a general way on a qualitative argument concerning the 'value' of the term in the retrieval context, or may involve a specific quantitative argument such as a proof of performance.

An example of the latter is provided by the relevance weighting theory [1]. Here it is proved that, under certain assumptions about term independence, optimum performance is achieved by using a simple sum-of-weights match function and giving a term t a weight:

$$w_t = \log \frac{p_t(1 - q_t)}{q_t(1 - p_t)}$$

where p_t is the probability that a given relevant document is assigned the term t, and q_t is the equivalent non-relevant probability (p and q may be estimated from relevance feedback information). The quantitative nature of the argument is well illustrated by the use of 'simple sum-of-weights' together with the logarithm in the formula: if one were to *multiply* the weights instead of adding them, then the same theory would demand that the logarithm was not used. A generalised qualitative argument about term value would not be capable of distinguishing the two cases.

Journal of Documentation, Vol. 46, No. 4, December 1990, pp. 359–364.

The argument below assumes a term weighting function similar to the above, though not necessarily this particular one.

Query expansion

Various methods have also been proposed for drawing in new terms to enhance a search statement. In a relevance feedback system, for example, terms may be drawn from items retrieved and judged relevant in a previous iteration.

Such query expansion may be automatic (i.e. the system finds and includes new terms without reference to the user), or semi-automatic (i.e. the system finds new terms and offers them to the user for possible inclusion). In the case of automatic query expansion in a weighted or associative retrieval system, it may be appropriate to throw in all candidate terms (e.g. all the terms in the known relevant documents), and leave the term weighting scheme to cope with the fact that some terms may be better than others. But this procedure may fall foul of the 'curse of dimensionality' [2].

For this reason, it may be necessary to have a term selection stage, so that apparently poor terms are not included at all, rather than being given low weights. In a semi-automatic system, it is necessary to present the terms to the user in some reasonable order, preferably one in which the terms most likely to be useful are near the top.

Either way, therefore, it seems appropriate to look for a measure of term value or usefulness, for the purpose of query expansion.

If, then, we have a term weighting scheme for retrieval, based on a general 'term value' argument, such an argument would presumably apply equally well for this new purpose. If, however, we have a more specific quantitative argument, it is not at all clear that the argument would apply to this new purpose as well as to the old. This suggests looking for a specific quantitative argument to apply to query expansion.

The object of this note, then, is to present a very rough sketch of such an argument, and to use the argument to suggest that an appropriate criterion for term selection is indeed different from an appropriate criterion for term weighting once selected.

2. SWETS MODEL

The argument presented here makes use of ideas taken from the Swets model of IR system performance [3]. The following is a brief description of the Swets model.

The system is assumed to retrieve items by ranking them according to some measure of association with the query (match function). The principal idea of the Swets theory is to examine the distribution of values of this match function over the document collection. More specifically, it considers two such distributions, one for the relevant documents and one for the non-relevant. If the retrieval system is any good, the two distributions will be different: in particular the match function values will generally be higher for relevant documents than for non-relevant. For example, the distributions might look like Figure 1. Here μ_R is the average (mean) match function value for relevant documents, and μ_N the mean for non-relevant documents.

In general, the more the two distributions are separated, the better the performance of the system will be. Other things being equal, the higher the difference $d = \mu_R - \mu_N$ between the means of the two distributions, the better the performance. Actually the measure of performance proposed by Swets, and an alternative proposed by Brookes [4], can both be expressed as d normalised by some function of the standard deviations of the distributions. However, these measures are associated with the assumption that the distributions are normal. This would not be an appropriate assumption for the present situation, as argued below. So the present argument is based on the use of d, unnormalised, as a simple measure of performance. For this reason and others, the argument is not regarded as a rigorous proof, but rather as indicative of relationships between the variables.

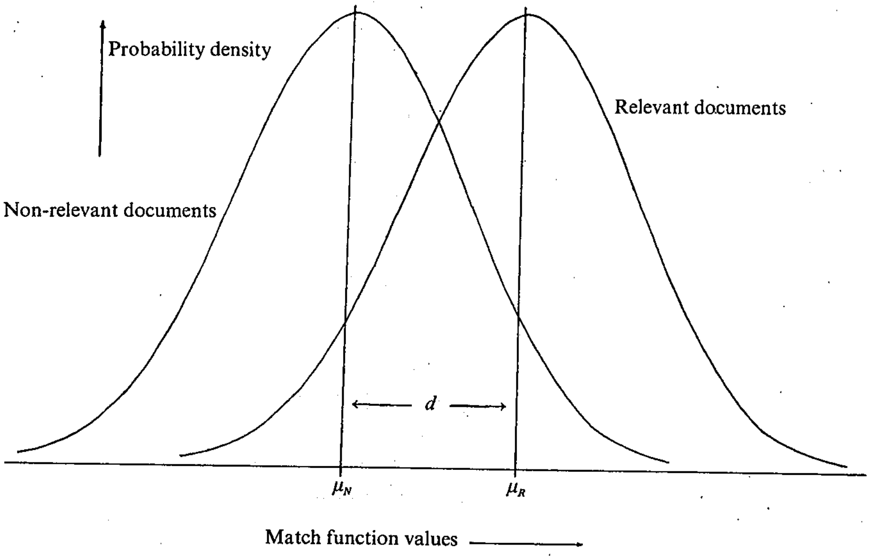

FIGURE 1. *The Swets model: distribution of match function values*

3. AN APPROACH TO TERM SELECTION

Assumptions

We suppose that we have an initial search formulation, using some match function, and that we are considering the possible inclusion of a new term, whose presence in an item would then add a certain quantity (i.e. the weight of the term) to the value of the match function. A simple sum-of-weights match function would satisfy this assumption, though it is not necessary to assume that the match function as applied to the initial search formulation is such a sum of weights. Nor is it necessary to assume that the weight in question is the relevance weight discussed above.

It is, however, necessary to make an additional strong simplifying assumption, again putting the argument in the realm of indication rather than proof. We assume statistical independence between the new term and the entire previous search formulation. This assumption can be expressed as follows.

Consider one of the two Swets distributions for the initial formulation, say the relevant one. The population of relevant items is further divided into those that contain the new term and those that do not, each of which has its own distribution of values for the match function applied to the initial formulation. The independence assumption, then, is that these two new distributions are identical, i.e. that the presence or absence of the new term does not affect the distribution of the old match function. This independence assumption must be applied separately to the non-relevant as well as to the relevant items.

There is no necessary contradiction between such independence assumptions and the idea of query expansion based on the results of an initial search. In fact the assumptions given predict a positive association between an initial query formulation and a good new term, when the whole collection is considered. In effect, it is assumed that such association can be completely explained in terms of relevance.

The assumptions are, nevertheless, strong ones. Some further discussions on similar independence assumptions may be found in [1].

How useful is the new term?

We now express the question, 'How useful would the candidate term be?' as 'How much effect would adding it to the search formulation have on retrieval performance?'

If the weight of the candidate term is w_t, then those items that contain the term will have w_t added to their match function values. The new Swets distribution for the relevant items (under the independence assumptions) consists of a mixture of the original and the original displaced upwards by w_t. Making use of the probability p_t defined in Section 1, the mixing proportions are $(1 - p_t):p_t$. Thus the new mean is:

$$(1 - p_t)\mu_R + p_t(\mu_R + w_t) = \mu_R + p_t w_t$$

Similarly, the new mean for the non-relevant items is:

$$\mu_N + q_t w_t$$

and the new effectiveness d' is:

$$\begin{aligned} d' &= \mu_R + p_t w_t - \mu_N - q_t w_t \\ &= \mu_R - \mu_N + w_t(p_t - q_t) \\ &= d + w_t(p_t - q_t) \end{aligned}$$

In other words, the inclusion of the term t in the search formulation, with weight w_t, will (under the assumptions given) increase the effectiveness by

$$a_t = w_t (p_t - q_t)$$

This suggests that, if we have a weighting function and estimates of p_t and q_t, new terms should be ranked in order of their a_t values for possible inclusion in the search formulation.

Notice that a_t is *not* w_t. In other words, if we have a good rule for assigning weights to terms once they are included in the search formulation, the same rule will *not* necessarily serve to decide whether to include the term in the first place.

Notice also that it is *not* a question of weighting new terms in a different way from old ones. Although no specific method is assumed here, the method of arriving at w_t can be (indeed probably should be) the same for both new and old terms, once the list of terms constituting the query is fixed. What the argument suggests is that the decision rule for including new terms (whether automatically or semi-automatically, as discussed above) should be based on a_t rather than w_t.

4. DISCUSSION

The result given in the previous section is, as argued above, to be treated as indicative rather than as formal proof. Nevertheless, it is clear that the two questions of term 'value' must be separated. and potentially at least have different answers. This statement is in part simply a recognition that the questions are different. It is worth trying to express them in a way which makes the difference clear:

1. Given that a term is included in a search formulation, how much evidence does it provide as to the relevance of an item in which it is present?
2. Given a candidate term, how much effect does its inclusion have on the overall effectiveness of the search formulation?

It now seems unsurprising that the answers may be different! It will also be clear that the nature of any formal approaches to answering the two questions should be different. In particular, the first question allows a stronger kind of optimisation than the second. The probabilistic model for search term weights attempts, via the Probability Ranking Principle [5], to optimise the entire length of the search curve from high-precision to high-recall. A model for which term(s) to add, however, will have to lead to a preference between a very infrequent term which contributes chiefly to the high-precision end, and a frequent term which contributes mainly to the high-recall end. The argument here must be based on some kind of average performance.

Future work

One would like to be able to tighten the formal argument and come up with a firmer indication of an appropriate term selection criterion. Unfortunately, the Swets model approach used here may not be suitable for tightening up.

One particular problem with the Swets model in this context lies in its explicit dependence on the match function values. A non-linear but monotonic transformation of these values will produce exactly the same retrieval effect but alter the effectiveness as measured in the Swets model. (An example of such a transformation would be the replacement of an additive function by a multiplicative one, as suggested in Section 1.) In the original applications of the Swets model, this possibility is effectively eliminated by the assumption that the distributions are normal. However, we cannot use that assumption here, since if the distributions were normal prior to adding the new term, they would not be afterwards.

It would therefore be desirable to develop a model that would not require explicit dependence on the match function values.

Clearly the present model (or any future model) needs also to be subjected to various kinds of empirical test. The object of this note has been only to present the theoretical argument; but no model in information retrieval can be accepted without empirical support.

Conclusion

A term weighting formula that provides appropriate weights for use in a match function for retrieval is not necessarily an appropriate measure for term selection in the first place.

Given a term weighting formula that provides weights w_t, and estimates of p_t and q_t (as defined in Section 1), a first suggestion for a measure for term selection is

$$a_t = w_t\,(p_t - q_t)$$

REFERENCES

1. ROBERTSON, S.E. *and* SPARCK JONES, K. Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 1976, 129–146.
2. VAN RIJSBERGEN, C.J. *Information retrieval.* 2nd ed. London: Butterworth, 1979.
3. SWETS, J.A. Information retrieval systems. *Science, 141*, 1963, 245–250.
4. BROOKES, B.C. The measure of information retrieval effectiveness proposed by Swets. *Journal of Documentation, 24*, 1968, 41–54.
5. ROBERTSON, S.E. The probability ranking principle in IR. *Journal of Documentation, 33*, 1977, 294–304.

(*Revised version received 5 April 1990*)