A comparison of the cosine correlation and the modified probabilistic model

W Bruce Croft

In a recent paper[1], we reported a series of retrieval experiments with search strategies based on a modified probabilistic model[2]. It has been pointed out that the comparison between the performance of the cosine correlation and the modified probabilistic model was incomplete. In particular, the term weights used for the cosine correlation were the term frequencies within the document text. Salton has for some time used a term weight known as '*tf.idf*' in his retrieval experiments with the cosine correlation[3]. This weight consists of the within-document term frequency (sometimes normalized by the maximum frequency) multiplied by the inverse document frequency weight[4]. Although the inverse document frequency weight can be regarded as a product of the *retrieval* process[5], it has also been used as part of the *indexing* process in that the weight is assigned to the terms in the document representatives. In this note, we shall present the results of retrieval experiments with the cosine correlation and the *tf.idf* weights. The comparison of these results to those obtained with the modified probabilistic model leads to some interesting conclusions about the cosine correlation.

In the following experimental results, WCOS refers to the cosine correlation used with the *tf.idf* weight where *tf* is the unnormalized within-document frequency. NCOS uses the cosine correlation/*tf.idf* weight where *tf* is the term frequency normalized by the maximum frequency in a document. NEW refers to the best results for the search strategy derived from the modified probabilistic model. The NEW strategy, when used before relevance feedback, assigns a score to a document according to the following expression

$$\Sigma\, ts_i . idf_i . q_i \qquad (1)$$

Computer and Information Science Department, University of Massachusetts, Amherst, MA 01003, USA
Received 14 February 1984

In this expression, ts_i is the term significance weight for the ith term in the current document, idf_i is the inverse document frequency weight for term i, and the q_i is the ith query term (q_i is assumed to be binary, either 0 or 1). The term significance weight, as described by Croft[1], is an estimate of the probability that a term is assigned to the current document representative. In that paper, the estimate used for this probability was $k + (1 - k)tf$, where tf is the normalized within-document frequency. The constant k (usually 0.5) was included to reflect the fact that if a term occurs in the document text at all, it should have a reasonably high probability of assignment.

The results are described in terms of the precision values at standard recall levels (0.1 to 1.0). Both the Cranfield and NPL collections were used.

Cranfield	Precision at Standard Recall									
WCOS	54.0	46.7	37.1	31.9	28.8	21.4	16.8	13.6	9.6	9.0
NCOS	53.2	46.4	36.9	31.6	28.4	21.4	16.7	13.6	9.5	9.0
NEW	53.8	47.4	40.2	35.3	31.9	23.1	17.6	14.1	10.2	9.6

NPL										
WCOS	40.3	31.9	25.4	20.4	15.3	11.3	8.7	6.6	4.2	2.2
NCOS	39.5	31.3	24.6	19.6	14.9	11.1	8.5	6.5	4.1	2.1
NEW	59.1	48.4	39.7	33.2	25.8	19.4	14.4	10.9	6.7	3.3

There are two main features of these results:

- Although the cosine/*tf.idf* search has similar results to NEW for the Cranfield collection, there is a large difference for the NPL collection.
- Normalizing the term frequencies in the *tf.idf* weight has no effect on performance.

The rest of this note provides an explanation for these results. The first part of this explanation is a transforma-

 0144-817X/84/020113-02$03.00

tion of the expression for the cosine correlation. This expression is

$$\Sigma t_i.q_i/\sqrt{\Sigma t_i^2 \Sigma q_i^2}$$

where t_i is the term weight for the ith term. This can be transformed into

$$(1/\sqrt{\Sigma q_i^2})\,\Sigma\,(t_i/\sqrt{\Sigma t_i^2}).q_i \qquad (2)$$

The first term in expression (2) is the same for all documents and, therefore, has no effect on the ranking. This expression shows that the cosine correlation is equivalent to a simple match where the term weights are normalized. If the term weight used is *tf.idf* and *tf* is not normalized, then the cosine correlation is equivalent to the modified probabilistic model (expression (1)) with the term significance probability estimated by a normalized within-document frequency. The normalization used (dividing by $\sqrt{\Sigma t_i^2}$) will give low estimates for these probabilities. Note that normalizing the *tf* component prior to searching should not have any effect because the cosine correlation already incorporates this normalization. This explains the similar results obtained with WCOS and NCOS.

The equivalence of the cosine/*tf.idf* search to a simple form of the modified probabilistic model also explains the difference in the results for the two collections. The experiments by Croft[1] show that, for the Cranfield collection, estimating the term significance by normalizing the within-document frequencies gave good results. However, for the NPL collection, simple normalization gave poor results and more sophisticated estimations were required to realize the benefits of the modified probabilistic model.

REFERENCES

1 **Croft, W B** 'Experiments with representation in a document retrieval system' *Inf. Technol.* Vol 2 No 1 (January 1983) pp 1–21

2 **Croft, W B** 'Document representation in probabilistic models of information retrieval' *J. Am. Soc. Inf. Sci.* Vol 32 (1981) pp 451–457

3 **Salton, G and Yang, C S** 'On the specification of term values in automatic indexing' *J. Doc. (GB)* Vol 29 (1973) pp 351–372

4 **Sparck Jones, K** 'A statistical interpretation of term specificity and its application to retrieval' *J. Doc. (GB)* Vol 28, (1972) pp 11–20

5 **Croft, W B and Harper, D J** 'Using probabilistic models of information retrieval without relevance information' *J. Doc. (GB)* Vol 35 (1979) pp 285–295