

THE PROBABILITY RANKING PRINCIPLE IN IR

S. E. ROBERTSON

*School of Library, Archive, and Information Studies,
University College London*

The principle that, for optimal retrieval, documents should be ranked in order of the probability of relevance or usefulness has been brought into question by Cooper. It is shown that the principle can be justified under certain assumptions, but that in cases where these assumptions do not hold, the principle is not valid. The major problem appears to lie in the way the principle considers each document independently of the rest. The nature of the information on the basis of which the system decides whether or not to retrieve the documents determines whether the document-by-document approach is valid.

A REFERENCE retrieval system should rank the references in the collection in order of their probability of relevance to the request, or of usefulness to the user, or of satisfying the user. This principle was first used explicitly by Maron and Kuhns.¹ Given that no system is capable of making a definitive assessment of relevance, it seems intuitively obvious that some such notion must be used; Maron and Kuhns accept the principle *a priori*. However, a closer analysis of the principle suggests that we need to examine carefully the assumptions on which it is based and the ways it might be interpreted. The object of this paper is to make a first attempt at such an analysis.

I. BACKGROUND

Maron and Kuhns's early paper introduced a very necessary new idea into discussion on the basic problems of retrieval. The idea was that since no retrieval system can be expected to predict *with certainty* which documents a requester might find useful, the system must necessarily be dealing with *probabilities*; we should therefore design our systems accordingly.

That said, the particular approach adopted by Maron and Kuhns in some ways confuses the issue (see Robertson).² They *define* the relevance of a document to an index term as the probability that a user using this term will be satisfied with this document: a definition which does not correspond to the usual use of the word 'relevance'.

In this paper, I will take relevance (or usefulness, or user satisfaction) to be a basic, dichotomous criterion variable, defined outside the system itself. The assumption of dichotomy is a strong one, and almost certainly not generally valid; discussions of a more complex model are given elsewhere.^{2,3} Several more possible assumptions about the nature of this criterion variable are discussed below.

Given a dichotomous criterion variable, and a system which has some (essentially probabilistic) information about this variable, it seems obvious enough that the documents which are most likely to satisfy the user should be presented to

him or her first. This idea has been used, in one form or another, by various people since Maron and Kuhns. Cooper⁴ gives a formal statement of the principle:

The probability ranking principle (PRP): If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.

However, Cooper goes on to show how counter-examples to the PRP can be constructed: situations in which a straightforward application of the principle leads to clearly less-than-optimal performance.*

Elsewhere,³ before making use of the PRP, I have given a formal justification for it, on the basis of certain assumptions. In the present paper, after a discussion of the nature of the criterion variable, I give two such justifications. In the remainder of the paper, I begin to explore the areas not covered by these assumptions, from which Cooper's examples are taken.

2. THE CRITERION VARIABLE

The object of a reference retrieval system is to predict, in response to a request, which documents the requester will find relevant to his request, or useful to him in his attempt to find the answer. Relevance or usefulness must thus be defined outside the system itself, as a criterion for the system. What assumptions can we make about the nature and characteristics of this criterion variable?

I have already indicated that I will assume the variable to be dichotomous: that is, a document is either relevant or not, there are no in-between states. Indeed, the very statement of the PRP implicitly makes the same assumption. We might consider a more general principle which would get over this problem: Cooper, for example, considers ranking by 'expected utility', where the utility of a document is a continuous variable. In this paper, however, I deal only with the dichotomous case.

Does the relevance or usefulness of one document affect the relevance or usefulness of another? There are several aspects to this question. First, the fact that document A has been retrieved before document B (= is higher up the ordered list presented to the user, and has presumably been seen first) may affect the usefulness of B, if for example B simply repeats the same information as A. Second, the fact that document A has already been judged relevant by the user may provide some indication of the possible relevance of B. Third, even if the system does not know whether A has been judged relevant, it may know that there is a correlation between the acceptance of A and B for different users. Fourth, two documents taken together may be relevant where neither one is relevant on its own if they each tackle complementary aspects of the problem; and so on.

All these possibilities may affect the use of the PRP in different ways. For the purpose of defining a case in which the PRP holds unequivocally I make the following simplifying assumptions:

* As Cooper's paper has not been published, I present his main counter-example in the Appendix to this paper.

- (a) The *relevance* of a document to a request is independent of the other documents in the collection;
- (b) The *usefulness* of a relevant document to a requester may depend on the number of relevant documents the requester has already seen (the more he has seen, the less useful a subsequent one may be).

These assumptions raise some interesting questions about the nature of the information the system might use to predict relevance, and about the nature of the probability of relevance. Such questions are discussed below; in the meantime, the assumptions form a suitable basis for a justification of the PRP.

3. FIRST JUSTIFICATION: TRADITIONAL MEASURES OF EFFECTIVENESS

The object of this section is to prove that, under certain conditions, the PRP leads to optimum performance, where performance is measured by means of parameters which are very close to the traditional measures of retrieval effectiveness. An earlier version of this proof was first presented elsewhere.³

We first demonstrate a general result concerning the probabilities of any two events a , b (\bar{a} denotes the event 'not a '). Two applications of Bayes's theorem give the following:

$$P(a|b) P(b) = P(a \cap b) = P(b|a) P(a)$$

Similarly:

$$P(\bar{a}|b) P(b) = P(b|\bar{a}) P(\bar{a})$$

Hence:

$$\frac{P(a|b)}{P(\bar{a}|b)} = \frac{P(b|a) P(a)}{P(b|\bar{a}) P(\bar{a})}$$

We now use the well-known logistic (or log-odds) transformation of a probability, which is defined by:

$$\text{logit } P(x) = \log \frac{P(x)}{1 - P(x)} = \log \frac{P(x)}{P(\bar{x})}$$

Hence:

$$\begin{aligned} \text{logit } P(a|b) &= \log \frac{P(b|a)}{P(b|\bar{a})} + \log \frac{P(a)}{P(\bar{a})} \\ &= \log \frac{P(b|a)}{P(b|\bar{a})} + \text{logit } P(a) \end{aligned}$$

So we have demonstrated the following:

Lemma: For any two events a , b ,

$$\text{logit } P(a|b) = \log \frac{P(b|a)}{P(b|\bar{a})} + \text{logit } P(a)$$

We now define the parameters of interest. The system is assumed to order (or partially order) the documents in response to a request, and a cut-off is applied to define the retrieved set. (Various possible algorithms which searchers might use to define the cut-off point are surveyed by Cooper;⁵ a few specific ones are considered below.) Our parameters are:

$$\begin{aligned} \theta_1 &= P(\text{document retrieved} | \text{document relevant}) \\ \theta_2 &= P(\text{document retrieved} | \text{document non-relevant}) \\ \phi &= P(\text{document relevant} | \text{document retrieved}) \\ \gamma &= P(\text{document relevant}) \end{aligned}$$

All these parameters relate to an individual request—indeed to an individual need. They correspond closely to the traditional proportion measures recall, fallout, precision and generality respectively (where these are calculated for individual requests); the exact relationship between the probabilities and the proportions can be expressed either as:

Recall is an *estimate* of θ_1

or as: θ_1 is *expected* recall

These matters are discussed further elsewhere.³

We also define some more parameters, relating to an individual document as well as to an individual request: for any given document d_i ,

$$\begin{aligned} \theta_1(d_i) &= P(\text{document is } d_i | \text{document relevant}) \\ \theta_2(d_i) &= P(\text{document is } d_i | \text{document not relevant}) \\ \phi(d_i) &= P(\text{document relevant} | \text{document is } d_i) = P(d_i \text{ is relevant}) \end{aligned}$$

Then we have:

$$\theta_1 = \sum_{d_i \in S} \theta_1(d_i)$$

$$\theta_2 = \sum_{d_i \in S} \theta_2(d_i)$$

(where S is the retrieved set)

Also $\phi(d_i)$ is the probability of relevance which the PRP says we should use to rank the documents.

From the lemma, we have:

$$\text{logit } \phi(d_i) = \log \frac{\theta_1(d_i)}{\theta_2(d_i)} + \text{logit } \gamma$$

or

$$\theta_1(d_i) = x_i \theta_2(d_i)$$

where x_i is monotonic with $\phi(d_i)$

(In particular, $x_i = \exp [\text{logit } \phi(d_i) - \text{logit } \gamma]$)

So if the cut-off is defined by a value of θ_2 , we should clearly optimize retrieval (maximize θ_1) by including in the retrieved set those documents with the highest values of x_i —that is, those with the highest values of $\phi(d_i)$. In other words maximum expected recall for given expected fallout is obtained by ranking in order of $\phi(d_i)$ and applying a cut-off when the given fallout is reached.

Similarly, we can show that expected fallout is minimized for given expected recall, or that expected recall is maximized and expected fallout minimized if a given number of documents is retrieved, using the same document ordering. We can also apply the lemma again:

$$\text{logit } \phi = \log \frac{\theta_1}{\theta_2} + \text{logit } \gamma$$

to show that expected precision is maximized under any of the three cut-off criteria mentioned. Further, we could extend the analysis to a number of other effectiveness measures that have been proposed, including for example Cooper's⁶ expected search length.

Thus we have proved that, under certain conditions, the PRP optimizes performance. With regard to the conditions, it should be noted that:

- (a) the entire formalism makes the assumption suggested in §2, that the relevance of a document to a request does not depend on the other documents in the collection;
- (b) the proof relates only to a single request; if a set of requests is being considered problems arise because the value of γ may differ from request to request, and because the measures of performance must in some way be averaged over the requests.

4. SECOND JUSTIFICATION: DECISION THEORY

The object of this section is to demonstrate that, under certain conditions, the PRP is the 'correct' decision procedure to use according to the dictates of Bayesian decision theory. An earlier version of this argument has been presented elsewhere.⁷

We define a 'loss function' associated with the decision as to whether or not to retrieve a document:

$$\text{Loss (retrieved|non-relevant)} = a_1$$

(that is, the loss associated with retrieving a non-relevant document is a_1), and

$$\text{Loss (not retrieved|relevant)} = a_2$$

Using the same notation as in the previous section, we suppose that we know the probability $\phi(d_i)$ of document d_i being relevant. If we retrieve it, the expected loss will be:

$$(1 - \phi(d_i))a_1$$

If we do not retrieve it, the expected loss will be:

$$\phi(d_i)a_2$$

So the optimum (loss-minimizing) decision is to retrieve d_i if:

$$\phi(d_i)a_2 > (1 - \phi(d_i))a_1$$

$$\text{or: } \frac{\phi(d_i)}{1 - \phi(d_i)} > \frac{a_1}{a_2}$$

$$\text{or: } \phi(d_i) > \frac{a_1}{a_2 + a_1}$$

Thus we can rank the documents in $\phi(d_i)$ order, and apply a cut-off where $\phi(d_i)$ falls below $a_1/(a_2 + a_1)$.

The assumption so far has been that a_1 and a_2 are constant for the situation under consideration. We can generalize the result somewhat, by supposing (as suggested in §2) that the usefulness of retrieving further relevant documents may diminish as some are retrieved. Then a_2 diminishes through the search; we have to recalcu-

late after each document is presented to the user. But we can still apply the same rule, stopping where $\phi(d_i)$ falls below the current value of $a_1/(a_2 + a_1)$.

Thus the PRP is valid in decision-theoretic terms, under certain conditions. Again it should be noted that:

- (a) the formalism again makes the assumptions about relevance and usefulness which were suggested in §2;
- (b) again there will be problems associated with applying the result to a set of questions, because the values of a_1 and a_2 assigned by the users may be different.

5. A DIFFERENT RANKING PRINCIPLE

Why does PRP fail? In the examples given by Cooper,⁴ the main problem seems to be in the calculation of the probabilities by cumulating over requests, where the requests vary in generality or total number of relevant documents. This clearly invalidates the argument of §3, which only works for a single request;* it also invalidates the decision theory approach of §4, since one cannot assume that the parameters a_1 and a_2 will be independent of the request.

In general, the PRP works document-by-document, whereas the results should be evaluated request-by-request. We can devise a form of ranking principle which works request-by-request which can be informally defined as follows:

Documents should be ranked in such a way that the probability of the user being satisfied by any given rank position is a maximum.

This alternative principle deals successfully with some of the situations in which the PRP fails, but there are many problems with it. In this section, I attempt to define its uses and limitations.

We have first to define what is meant by 'satisfaction'. We must assume that the user searches down the ranked list until he is in some sense satisfied with the information obtained; we have to exclude, for example, the possibility of a 'frustration-point' cut-off criterion (in Cooper's⁵ terms). We must also assume that this satisfaction is dichotomous. Examples of this kind of criterion are that the user is satisfied when he has retrieved a certain number or proportion of the relevant documents (such criteria are used by Keen and Digger⁸ in their experiments).

The second problem with the request-based principle is exactly that it does not work document-by-document; it is difficult to imagine an algorithm which would find the correct order in any situation, other than the ridiculously clumsy method of looking at all possible rankings.† The third problem arises from the second: there may not exist an optimal ranking in the terms of the principle, because maximizing the probability of satisfaction at rank 1 may mean excluding

* One could generalize the argument of §3, by defining all the probabilities in terms of the set of requests rather than a single request. However, this would involve accepting as measures of effectiveness the 'micro-average' values of recall etc.—that is, the ratios of cumulative numbers of documents—for each document cut-off level. Such measures are in general of dubious value, and are clearly inadequate for the situation of Cooper's example, where different users will stop searching at different points.

† Stirling¹⁰ has analysed the problem of devising such an algorithm. It is possible to do so, but the algorithm is still very time-consuming, and probably unusable for large collections, Stirling has demonstrated that it does, indeed, work better than the PRP.

the maximal probability of satisfaction by rank 10. We can demonstrate this fact with a simple example (a modification of one of Cooper's examples⁴), as follows.

We suppose that the system receives a request which indicates that the requester might fall into one of two groups: that is there are two distinct need-groups whose need is represented by the same formalized request. If he belongs to the first group (assume probability $2/3$), there are three documents (d_1-d_3) which will interest him, and he wants to see all of them; if he belongs to the second, there is just one document (d_4) which he wants. The two obvious rankings in which the system could present the documents are:

Ranking A: d_1, d_2, d_3, d_4
Ranking B: d_4, d_1, d_2, d_3

If we plot the probability of satisfaction against rank for these two rankings, we obtain the graphs shown in Fig. 1.

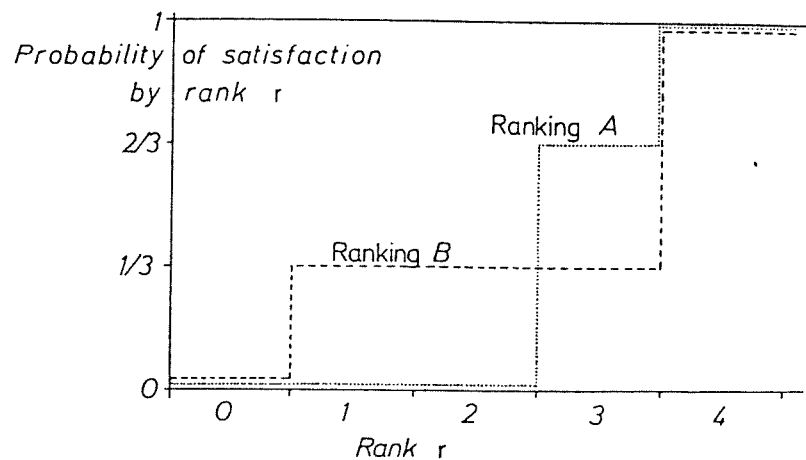


FIG. 1. Probability of satisfaction at each rank

Thus neither ranking maximizes the probability of satisfaction at every rank—indeed, it is not possible to do so.

We might therefore consider some modification of the principle to deal with this problem—say by maximizing the area under the curve of Fig. 1, or equivalently minimizing the area above the curve and bounded by the lines shown is the expected search length or rather its arithmetic mean over the requests.

It can also be related to recall and fallout: we would be minimizing fallout (or rather, the micro-average of fallout over requests) for the given recall cut-off. Thus we end up with the following possible principle:

Documents should be ranked in such a way as to maximize the area under the curve of probability of satisfaction against rank; or, equivalently, in such a way as to minimize expected search length (averaged over requests); or, equivalently, in such a way as to minimize fallout (micro-averaged over requests).

Such a principle, it should be noted, makes some assumptions about relative cost of (or losses due to) different events. For example, it is assumed equally valuable to reduce a comprehensive search from 100 to 99 documents, as it is to reduce a quick reference search from three to two documents. This assumption may be justified in cost-to-the-organization terms, but is unlikely to reflect the user-perceived value of the system. Unfortunately, the dilemma is not resolvable: in some circumstances an optimal ranking under one criterion *cannot* be optimal under another criterion.

The above principle is not likely to be of much practical use, for the two reasons given earlier:

- the fact that it relates only to satisfaction-point cut-off searches;
- the difficulty of devising an efficient algorithm for implementing it.

However, in the next section I show how the ideas behind it can shed some light on the original PRP.

6. DOCUMENT-BY-DOCUMENT APPROACHES

Ideally, one would like a ranking principle which, like the PRP, depends on the calculation of a single figure measure for each document, and the ranking of the documents in the order determined by this measure. However, in view of the discussion in the previous section, it seems unlikely that such a principle exists which could optimize performance under a range of conditions.

Is there any way in which we can modify the PRP in order to bring it closer to the request-based principle described in the last section, while maintaining its essentially document-by-document character? Two means suggest themselves.

We consider first the relationship between different need-groups represented by identical formalized requests, as in the example of the previous section. In that example, the PRP failed to give the ranking suggested by the request-based principle because it treated all relevant documents the same, even though one need-group required three documents for satisfaction, while the other group required only one. If we define the utility of a relevant document to be such that a total of one unit of utility is required for satisfaction, then we can rank the documents in order of expected utility. Thus in the example, documents d_1-d_3 are assumed to have utility $1/3$ for the first need-group; document d_4 has utility 1 for the second need-group. Expected utility then yields the correct ranking according to the request-based ranking principle.

The second modification of the PRP relates to another example of Cooper's. We now assume the same situation as in the previous example, except that the first need-group is satisfied by any *one* of the documents d_1-d_3 . The usual PRP suggests ranking the documents:

$$d_1, d_2, d_3, d_4$$

whereas the request-based principle suggests:

$$d_1, d_4, d_2, d_3$$

A modification of the PRP which gives the 'right' answer involves recalculating the probabilities of usefulness of the remaining documents after each document is retrieved. This modification, however, already destroys part of the document-by-document character of the PRP.

These two modifications do not by any means completely solve all the problems: one can make up more complex examples which show that the PRP, even with these two modifications, does not always produce the optimum ranking. They do, however, go some way towards bridging and interpreting the gap between the simply stated PRP and the general problem of finding a ranking which optimizes performance.

7. PROBABILITY AND INFORMATION

I now return to the basic idea behind the PRP. At the beginning of §2, I defined the object of a reference retrieval system as being to predict, in response to a request, which documents the requester will find relevant or useful. It is precisely this prediction process which the PRP was intended to formalize.

However, the discussions in Cooper's⁴ paper and above suggest that the estimation of a 'probability of relevance' for each document may not be the most appropriate form of prediction. Going back to the definition, we can identify two main questions:

1. On the basis of what kinds of information can the system make the prediction?
2. How should the system utilize and combine these various kinds of information?

These questions represent, indeed, the central problem of retrieval theory. All theories or hypotheses about information retrieval (in the sense of reference or document retrieval) relate ultimately to these two questions, and all retrieval systems are based on some (explicit or implicit) theory or hypothesis.

As an example of a hypothesis which relates to the two questions, consider the idea of clustering documents. The justification for clustering documents has been expressed by van Rijsbergen and Sparck Jones⁹ as the Cluster Hypothesis: that relevant documents are more like one another than they are like non-relevant documents. To express this hypothesis in prediction terms, the idea is that if document A matches the request and document B looks like document A, then this tells us something about the probable relevance of B, whether or not B itself matches the request.

What makes this example particularly interesting is that it is clear that the Cluster Hypothesis *cannot* be incorporated directly into a document-by-document calculation of probability of relevance, since the probability of relevance of document B in the example depends on the presence in the collection of document A. In practice, document cluster theory attempts to reduce the problem to the usual document-by-document form by first detecting the clusters, and then adding the information about cluster membership to the individual document records in some way. The subsequent retrieval operations can then be conducted in a document-by-document fashion.

No cluster theorist to my knowledge has attempted to approach the whole problem in terms of prediction. However, Goffman¹¹ suggested a method of retrieval based on the dependence between documents, and Croft and van Rijsbergen¹² have recently pointed out the similarities between Goffman's method and the clustering approach (and have given experimental evidence that the two methods give similar performance).

8. CONCLUSIONS

A retrieval system has to predict the relevance of documents to users on the basis of certain information. Whether the calculation of a probability of relevance, document-by-document, is an appropriate way to make the prediction depends on the nature of the relevance variable and of the information about it. In particular, the probability-ranking approach depends on the assumption that the relevance of one document to a request is independent of the other documents in the collection.

This assumption is as much about the nature of the information on the basis of which the system is trying to predict relevance, as it is about the nature of relevance itself. There are various kinds of dependency between documents, at various levels between the relevance itself and the information about it.

The probability ranking principle and its application can be regarded as a general theory of document-by-document information retrieval. But there exists no comparable theory for dealing with the more general problem of how to take account of dependency information. While dependency-oriented approaches to information retrieval, such as cluster-based retrieval, continue to be proposed, the development of such a general theory would seem to be of high priority.

REFERENCES

1. MARON, M. E. and KUHNS, J. L. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7, 1960, 216-44.
2. ROBERTSON, S. E. The probabilistic character of relevance. *Information Processing and Management*, 13, 1971, 247-51.
3. ROBERTSON, S. E. *A theoretical model of the retrieval characteristics of information retrieval systems*. Ph.D. thesis, University of London, 1976.
4. COOPER, W. S. The suboptimality of retrieval rankings based on probability of usefulness. (Private communication.)
5. COOPER, W. S. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24, 1973, 87-100 and 413-24.
6. COOPER, W. S. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19, 1968, 30-41.
7. ROBERTSON, S. E. and SPARCK JONES, K. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 1976, 129-46.
8. KEEN, E. M. and DIGGER, J. A. *Report of an information science index languages test*. College of Librarianship Wales, Aberystwyth, 1972.
9. VAN RIJSBERGEN, C. J. and SPARCK JONES, K. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29, 1973, 251-7.
10. STIRLING, K. H. The effect of document ranking on retrieval system performance: A search for an optimal ranking rule. *Proceedings of the American Society for Information Science*, 12, 1975, 105-6.
11. GOFFMAN, W. An indirect method of information retrieval. *Information Storage and Retrieval*, 4, 1969, 361-73.
12. CROFT, W. B. and VAN RIJSBERGEN, C. J. An evaluation of Goffman's indirect retrieval method. *Information Processing and Management*, 12, 1976, 327-31.

APPENDIX

The purpose of the appendix is to present in summary the counter-example to the probability ranking principle discovered by W. S. Cooper.⁴

Cooper considers the problem of ranking the output of a system in response to a given request. Thus he is concerned with the class of users who put the same request to the system,

and with a ranking of the documents in response to this one request which will optimize performance for this class of users.

Consider, then, the following situation. The class of users (associated with this one request) consists of two sub-classes, U_1 and U_2 ; U_1 has twice as many members as U_2 . Any user from U_1 would be satisfied with any one of the documents D_1 - D_9 , but with no others. Any user from U_2 would be satisfied with document D_{10} , but with no others.

Hence: any document from D_1 - D_9 , considered on its own, has a probability of $\frac{2}{3}$ of satisfying the next user who puts this request to the system. D_{10} has a probability of $\frac{1}{3}$ of satisfying him/her; all other documents have probability zero. The probability ranking principle therefore says that D_1 - D_9 should be given joint rank 1, D_{10} rank 2, and all others rank 3.

But this means that while U_1 users are satisfied with the first document they receive, U_2 users have to reject nine documents before they reach the one they want. One could readily improve on the probability ranking, by giving D_1 (say) rank 1, D_{10} rank 2, and D_2 - D_9 and all others rank 3. Then U_1 users are still satisfied with the first document, but U_2 users are now satisfied with the second. Thus the ranking specified by the probability-ranking principle is not optimal. Such is Cooper's counter-example.

It might be argued that in this particular situation one could do something different anyway: for example, U_1 and U_2 users might distinguish between themselves, given a suitable prompting device such as a 'see also' heading. But the basic point remains: given some data about the possible relevance or usefulness of certain documents to certain people, a strict application of the probability ranking principle does not necessarily lead to optimum performance.

The above counter-example is the basis for the example used in §6 of the present paper. That of §5 is a second example presented by Cooper.

(Received 12 July 1977)