THE

Journal of Documentation

VOLUME 34 NUMBER 2 JUNE 1978

RANKING IN PRINCIPLE

S. E. ROBERTSON

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

and

N. J. BELKIN

Centre for Information Science, The City University, London

It is often suggested that information retrieval systems should rank documents rather than simply retrieving a set. Two separate reasons are adduced for this: that relevance itself is a multi-valued or continuous variable; and that retrieval is an essentially approximate process. These two reasons lead to different ranking principles, one according to degree of relevance, the other according to probability of relevance. This paper explores the possibility of combining the two principles, but concludes that while neither is adequate alone, nor can any single all-embracing ranking principle be constructed to replace the two. The only general solution to the problem would be to find an optimal ranking by exploring the effect on the user of *every* possible ranking. However, some more practical approximate solutions appear possible.

INTRODUCTION

INFORMATION RETRIEVAL (IR) is concerned with the following situation:

(*a*) A user, recognizing an information need, presents to an IR system (i.e. a collection of texts, with a set of associated activities and mechanisms) a request, based upon that need, hoping that the IR system will be able to satisfy the need.

(*b*) The task of the IR system is to present the user with the text (or texts) which it judges to be most likely to satisfy the user's information need, based upon the request put to the system.

(*c*) The user examines the text, or some or all of the texts, presented by the system, and her/his need is satisfied completely or partially or not at all. The user's judgement as to the contribution of each text in satisfying the need establishes the *usefulness* or *relevance* of that text to the need.

The manner in which the IR system presents the user with the text(s) is a matter of interest. In particular, the system may present only one text, or an unordered set, or may rank the texts in some order. But any one of these processes (or indeed,

Journal of Documentation, Vol. 34, No. 2, June 1978, pp. 93–100.

any act of retrieval) *is* a ranking process (see Cooper[1]). For instance, in choosing a set of texts which are judged to be relevant to a request (standard set retrieval), a two-position ranking has been placed on the entire collection of texts.

An explicit ranking, beyond simple set retrieval, is normally based on a matching process: it reflects the *degree of match*, as measured by a *matching function*, between the texts and the request as put to the system. A number of systems have been designed to incorporate such a matching and ranking mechanism (e.g. Salton);[2] also, many theorists have considered such ranking to be an essential part of the IR process (e.g. Stokolova).[3] Even simple set retrieval can be regarded in the same light: the matching function simply has two values, 'Yes' and 'No'.

What reason is there for having a particular rank order? That is, why *should* one text be presented to the user before another? There are two quite distinct answers to this question (see Belkin),[4] which have been used by different people at different times, and have sometimes been confused. Since these two reasons or ranking principles might well generate different ranking rules or matching functions, and a combination of the two might lead to something different again, it is desirable that the two should be elucidated. This paper is an attempt to do just that.

TWO REASONS FOR RANKING

The first reason for ranking the texts is that different texts may satisfy the information need to different *degrees*. A formal model for relevance as a variable of degree is described by Robertson.[5] It is assumed that underlying any statement of relevance is a continuous scale on which individual texts have positions; one end of the scale represents high relevance or pertinence or usefulness or satisfaction, the other indicates non-relevance, non-pertinence, uselessness, or dissatisfaction.*

Any use of such a model as justification for ranking texts involves an additional assumption. One must assume that the matching function *is* in some sense a measure of this continuous variable underlying relevance. Thus, for example, Radecki[6] (pp. 319–20) says: 'the concept of relevance is expressed by the . . . degree of similarity between the search patterns of documents and the search patterns of the information requests.'

The second reason for wanting to rank texts is based, by contrast, on the more primitive notion of relevance as a strictly dichotomous variable: a text either is relevant to a need or is not. Under this assumption, the ideal response of an IR system is not a ranked list: it is to retrieve exactly that *set* of texts which will be judged relevant. But of course the system cannot do this; it cannot be certain that any given text will be judged relevant, it can only make a probabilistic statement about relevance. In these circumstances, it seems clear that the system should respond by ranking the texts in order of their *probability of relevance* (according to the information available to the system). That is, the information is incorporated into a matching function, degree of match being interpreted as probability of relevance.

Much previous work in IR has involved setting up a system (or defining a theoretical system) with a plausible matching function, and then claiming one or both of the above arguments as justification for use of the matching function. In the process, the two arguments are often confused.

* Hereafter we use just the word *relevance* to refer to this entire group of concepts, unless there is some reason to specify one in particular.

EXAMPLES OF PREVIOUS WORK

A recent contribution by Stokolova[3] is an example of the problems which arise, and is by no means unique in this respect. She begins by examining the nature of the relevance relationship between text and request, noting that the 'subject experts' who make the necessary judgement '... recognize not only the presence or absence of relevance between such pairs but also indicate different degrees (i.e. some measure) of relevance. The output of many IR systems is ranked in order to take account of these differences.' (Stokolova,[3] p. 227). Here one must assume that Stokolova has adopted the *degree of relevance* position, and that her ranking mechanism will be one based on this principle. Yet we find that when it comes to actually formalizing her ranking rule, she switches from one assumption to the other: 'It has already been mentioned that such subject experts are aware of *different degrees of relevance* of documents to request, recognizing that documents may have *different probabilities of* "answering" a request' (Stokolova,[3] p. 229; emphasis ours.) Thus, her ranking mechanism is according to the (supposed) *probability* that a document in the collection will be 'strictly relevant' (Cooper)[7] rather than to the *degree* to which a document is relevant to a request. Although Stokolova has (apparently implicitly) recognized the two assumptions on which ranking in IR is based, she has been unable to use them both (or has confused them with one another) in her own work.

Others have exhibited similar confusion, or lack of discrimination. Thus, for example, Maron and Kuhns[8] have a 'relevance number' which is defined as a probability of relevance or satisfaction. Brandhorst and Eckert[9] (p. 402) say 'weighted term searches can provide a "rating of probable pertinence".' Stiles[10] sums weights to provide a 'document relevance number', which Doyle[11] (p. 273) describes as a 'measure of probable document relevance'. Gebhardt[12] defines the 'relevance value' of a document as its expected (mean) value on a multi-valued relevance scale, when judged by different judges. In these cases and many others, a theoretical IR system is described in which both ranking principles appear to be simultaneously invoked, although rarely explicitly.

There is, however, a marked asymmetry in the further analysis and use of these two distinct principles. The principle that documents should be ranked by probability of (dichotomous) relevance has been the subject of some analysis (see Robertson),[13] and has also been used directly to determine the choice of a matching function (e.g. Robertson and Sparck Jones).[14] The principle of ranking according to degree of relevance, by contrast, has not yet (to our knowledge) had any direct effect on the choice of matching function. If it were to do so, there is no reason to assume that a matching function dictated by one principle would be the same as one dictated by the other.

The problems become apparent again in evaluation experiments on IR systems. Although most evaluators ask their judges to provide relevance assessments on a scale with three or more levels, they then seem to have little idea how these scales can be used in the evaluation. Usually, they reduce the multi-valued scale to a dichotomous one, by choosing a threshold or cut-off level, and make all subsequent manipulations such as the calculation of measures of performance in terms of the dichotomous scale. In effect, the implicit ideal of these tests is a ranking according to probability rather than according to degree.

On the other hand, Lesk and Salton,[15] in an investigation of the effect of different judges of relevance on the evaluation process, show experimentally that high-relevance documents (on which judges are most likely to agree) do tend to be ranked higher by IR systems than fringe documents (on which judges often disagree). But this is only a qualitative result; there appear to be no methods in use at present which can evaluate the ability of a system to rank in order of degree of relevance.

It is clear from the examples given that, on the one hand, there is a general recognition that relevance should be regarded as a variable of degree rather than a dichotomous property, but there is little agreement on how this fact should be used or taken account of. On the other hand, there is also a recognition that the statements that a system can make about relevance are essentially probabilistic ones; but all applications of *this* fact to date have involved the assumption that relevance is dichotomous. Thus, although theoretical systems often employ the degree of relevance principle (usually implicitly), by the time the system is constructed or evaluated, this aspect of ranking has unobtrusively disappeared.

DISCUSSION OF THE PRINCIPLES

Given that there appears to be an intuitive understanding of the two ranking principles which we have discussed, why has the confusion of these ranking principles arisen, and why have experimenters and evaluators been unable to use both? It seems to us that the fundamental problem lies in some implications of the IR situation as described above, in particular in the relationships between request and need and between text and need.

To understand these implications, we first have to analyse the two ranking principles in terms of the description of the IR situation. *Degree* of relevance has nothing directly to do with the request (formal statement of the need), nor with the IR mechanism, but rather strictly with the relationship between text and need, in the judgement of the user (stage 'c' of the IR situation as described above). It is hypothesized that the user will judge some texts to be more relevant to (or more useful in satisfying) the need than others. *Probability* of relevance, by contrast, has to do with the discrepancy between request and need (stage 'a'), and with the system's ability (stage 'b') to make inferences about the need on the basis of the request. If the user could state her/his need completely and exactly, and if the indexing of the texts were also complete and exact, then the probability concept would not arise: perfect retrieval would be possible.

How do the two principles stand up to somewhat deeper analysis? First it should be pointed out that there are situations in which neither applies; that is, where the question is either answered or not, and where the need is (or could in principle be) stated completely and exactly. These are the situations to which Cooper's[7] notion of 'logical relevance' applies, and where data or fact retrieval systems are appropriate. More generally, though, we take it as axiomatic that there will be discrepancies between requests and needs, and thus that some probabilistic ideas must enter into retrieval.

Even making this assumption, and assuming dichotomous relevance, the probability ranking principle does not automatically follow. It has been shown (see Cooper;[16] Stirling;[17] Robertson[18]) that the probability ranking principle can lead to non-optimum results, and that a replacement algorithm which does not suffer these defects *cannot be* stated simply as a ranking rule.

When we come to degree of relevance, however, the concept itself presents serious problems. Although it seems reasonable on an intuitive level to suppose that some texts will satisfy an information need more than others, it is not at all clear how this idea would fit in with any formal notion of relevance. For example, any direct extension of Cooper's logical relevance might not accommodate such a variable (e.g. Stokolova[8]); but Cook's[18] model of relevance-related behaviour, more akin to Cooper's[19] *utility*, does have an explicit continuous-relevance variable which is a reduction of a multi-dimensional assessment of the documents. Belkin's[20] *anomalous state of knowledge* model, while not explicitly containing such a variable, suggests that the user will judge the texts according to the extent to which they succeed in resolving the anomaly; again, the idea is perhaps more akin to Cooper's utility than to traditional notions of relevance.

However, it is not our aim here to produce or analyse any model of relevance. We take it that, given the IR situation as described, one must at least consider relevance as a continuous, rather than dichotomous variable, even if subsequent work on relevance indicates that the picture is more complex still.

COMBINING THE PRINCIPLES

Given then a continuous relevance variable (reflecting the discrepancy between text and need) and the discrepancy between request and need, one comes to the conclusion that the IR system must make some probabilistic statement about each text's degree of relevance to the need, if it is to treat the IR problem as a whole. That is, the two ranking principles should somehow be combined to produce a document ranking. Previous work appears to have failed at precisely this point; that is, in the examples we have discussed, the two principles, although often confused, have not been combined in any way that will result in a ranking which takes both variables into account. And herein lie the central problems of IR: just how can the concepts of probability and degree of relevance be combined in a statement which can predict the effect of a text on a user's need; and how should such a statement be used to provide an optimal ranking for retrieval?

We do not have any general solutions for either of these problems. That the first is non-trivial is clear from previous experience, although we suppose that there might exist some appropriate means for its solution. However, we hope to show, through an example, that given the two variables of which we must take account, the second problem is also non-trivial, and may perhaps be intractable, even if the first is resolved. That is, there is no obvious single ranking principle to replace the two discussed above.

The example concerns the question, 'Does cigarette smoking cause lung cancer?' We suppose that there are no documents in the IR system which deal directly and explicitly with this question, but there are two documents which must be considered. The first shows that there is some correlation between smoking and lung cancer, either between individuals or between geographical areas: this is obviously of some relevance (though it does not actually answer the question). The second shows that some other material or process, say X, causes lung cancer, but the system does not know of any relation between X and cigarette smoking. X could be, for example, nicotine or tar (in which case the document is probably highly relevant). It could be some environmental condition, such as level of industrial development, which provides an alternative explanation of a geographical correlation (in which case the document probably has some relevance). Or it could be something completely unrelated, such as aerosol car paint or asbestos dust (in which case the document is probably not relevant at all).

Given the information available to it, which of the two documents should the system retrieve first? The answer is not obvious: the simultaneous presence of the two variables, degree and probability, eliminates the obvious ranking principles or scales.

What this example seems to suggest is that we cannot hope to express the function of the IR mechanism in terms of a single all-embracing ranking principle; in particular, neither of the two principles discussed above, nor some simple combination, appears sufficient.

THE FUNCTION OF THE IR MECHANISM

Is there some way to move beyond this somewhat dispiriting conclusion? We suggest that an answer may lie in redefining the function of the IR system in terms of its task; that is, satisfaction of the user's information need. The two ranking principles have assumed that ranking by probability *or* by degree of relevance is the answer to this task, and the ranking principle is then taken to be the function of the system. Either of these principles conveniently translates into a matching function; that is, a single, separately measurable characteristic of each text which can be used to determine the rank order. Thus, in either case, the solution of the first part of the IR problem (that of making a predictive statement about relevance) suggests an obvious solution of the second part (that of using this statement to optimize retrieval). But we have seen that we should take account of both variables in making a predictive statement about relevance; and our example shows that such a statement will not translate so readily into a ranking mechanism; that is, its use in retrieval will not be so obvious.

Can we conceive of a *general* solution to the second part of the IR problem? Given that some form of ranking must be involved, what we require in general is a ranking of the texts that optimally satisfies the user's need. The general solution can only be to look at the effect on the user of every possible ranking of the texts. Such a procedure is clearly out of the question in practical terms. But Cooper[16] and Stirling[17] are forced to use a procedure that is almost as elaborate, even under an assumption of dichotomous, relevance in cases where the straightforward probability ranking principle falls down.

In any case, it is clear that the solution of the second part of the IR problem depends on what kinds of solutions to the first part are possible. This in turn depends on a fuller understanding of how users use texts in order to resolve their information needs. We suggest that further investigation and/or modelling of this process is vital for advance in solving these problems.

For example, a very simple model may provide an adequate (if imperfect) solution. If we assume that the texts have some simple cumulative effect, measured by a utility measure (in strictly utility-theoretic terms), then ranking by expected utility may be appropriate. However, we should again stress that such a model may not lead to a single ranking rule: as was seen above, even under the assumption of dichotomous relevance, the probability ranking principle is not always strictly optimal.

Perhaps more important, the model highlights the sort of strong assumptions about the resolution of user need that are commonly made. The assumption that the effect of successive texts is simply cumulative (in a quantitatively measurable way) is clearly far from the truth. Further research on the processes which users go through in resolving their needs may help us to develop a more adequate model. We make these comments not to suggest that these are the only routes to a solution of the problem, but rather to indicate the difficulties that are raised, even simply in defining the function of an IR system, by analysing and recognizing the complexity of the IR situation, and to identify the real extent of the IR problem.

CONCLUSION

We began this paper intending to show that there exist at least two principles of text ranking in IR, neither of which is sufficient in itself for dealing with the IR situation. In the course of the analysis of the IR situation, it has become clear that some combination of these two principles will be necessary for a satisfactory ranking principle to be constructed. Yet no obvious combined principle presents itself.

The possibility thus exists that the idea of a single ranking principle for IR is illusory, and that the search for such a principle is misguided. However, before giving up on this approach altogether, we suggest that solutions may lie in the relationships between text and user's need. More research is required on the processes which the user goes through in satisfying her/his need, and on developing adequate models of these processes and relationships. It is possible that such research will only confirm the tentative conclusion reached here, but it seems certain that the only means to developing a single ranking principle for IR, if one is possible, is through such models.

REFERENCES

1. COOPER, W. S. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation, 9*, 1968, 30–41.
2. SALTON, G. *The SMART retrieval system—experiments in automatic document processing.* Englewood Cliffs, N.J.: Prentice Hall, 1971.
3. STOKOLOVA, N. A. Elements of a semantic theory of information retrieval: I. The concepts of relevance and information language. *Information Processing and Management, 13*, 1977, 227–34.
4. BELKIN, N. J. The problem of matching in information retrieval. Paper presented at SIRE, Second International Research Forum in Information Science, Copenhagen, August 1977. To be published in *Information Reports and Bibliographies.*
5. ROBERTSON, S. E. The probabilistic character of relevance. *Information Processing and Management 13*, 1977, 247–51
6. RADECKI, T. New approach to the problem of information system effectiveness evaluation. *Information Processing and Management, 12*, 1976, 319–26.
7. COOPER, W. S. A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*, 1971, 19–37.
8. MARON, M. E. *and* KUHNS, J. L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM, 7*, 1960, 216–44.
9. BRANDHORST, W. T. *and* ECKERT, P. F. Document retrieval and dissemination systems. *Annual Review of Information Science and Technology*, 7, 1972, 379–438.
10. STILES, H. E. The association factor in information retrieval. *Journal of the ACM, 8*, 1961, 271–9.
11. DOYLE, L. B. *Information retrieval and processing.* Los Angeles: Melville, 1975.
12. GEBHARDT, F. A simple probabilistic model for the relevance assessment of documents. *Information Processing and Management, 11*, 1975, 59–65.
13. ROBERTSON, S. E. The probability ranking principle in IR. *Journal of Documentation, 33*, 1977, 294–304.
14. ROBERTSON, S. E. *and* SPARCK JONES, K. Relevance weighting of search terms. *Journal of the ASIS, 27*, 1976, 129–46.
15. LESK, M. E. *and* SALTON, G. Relevance assessments and retrieval system evaluation. *In:* Salton,[2] pp. 506–27.

16. COOPER, W. S. The suboptimality of retrieval rankings based on probability of usefulness. (unpublished).
17. STIRLING, K. H. The effect of document ranking on retrieval system performance: a search for an optimum ranking rule. *Proceedings of the ASIS, 12*, 1975, 105–6.
18. COOK, K. H. A threshold model of relevance decisions. *Information Processing and Management, 11*, 1975, 125–35.
19. COOPER, W. S. On selecting a measure of retrieval performance. *Journal of the ASIS, 24*, 1973, 87–100, 413–24.
20. BELKIN, N. J. *A concept of information for information science*. Ph.D. Thesis, University of London, 1977.

(*Received 21 November 1977*)