Information Processing & Management, Vol. 13, pp. 247–251. Pergamon Press 1977. Printed in Great Britain

THE PROBABILISTIC CHARACTER OF RELEVANCE

S. E. ROBERTSON
School of Library, Achive and Information Studies, University College London, Gower Street, London WC1E 6BT, England

(*Received* 5 *January* 1976)

Abstract—Gebhardt's[1] probabilistic model of relevance is examined and found not to represent adequately some characteristics of the relevance judgement process. An alternative model is proposed, which identifies two different types of "error" or probabilistic variation between relevance judgements. The two types arise from, first, the definition of the boundaries of the relevance classes, and secondly the actual assessment of an individual document on the underlying scale (which is assumed to be a continuum). The problems of quantifying the model, and of assessing its implications for retrieval testing, are discussed.

A recent paper by GEBHARDT[1] considers the problem of setting up a probabilistic model to describe the way in which different jurors make different judgements of the relevance of a given document to a given problem. In principle, the idea is a good one and should serve to improve our understanding of the results of retrieval tests. But in practice, Gebhardt appears to make his probabilistic model too simple, and thus his specific conclusions are dubious. Further, Gebhardt does not consider or mention any of the projects investigating relevance which took place in the 1960s, whose results could have a direct bearing on his model. The two major projects are of course those undertaken by SDC (CUADRA and KATTER[2]) and by Case Western (REES and SCHULTZ[3]).

The purpose of the present paper is to propose a somewhat more complex, but hopefully very much more realistic, model of the relevance judgement process. I also hope to begin to relate the model to the results of experimental studies of relevance, as reviewed by SARACEVIC[4].† Considerable development work is required on the model before we can see exactly what the implications are for retrieval experiments; but I think it important that Gebhardt's conclusions should not be accepted by default at this stage.

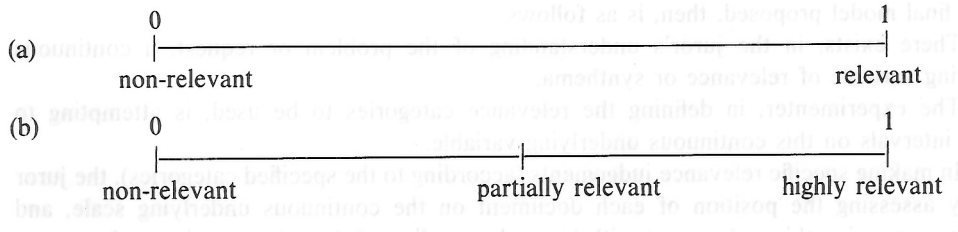
RELEVANCE AND THE JUDGEMENT PROCESS

The first conclusion that Gebhardt draws concerns the question of whether relevance should be judged in a binary manner (relevant/non-relevant) or on a more extensive scale. Unfortunately, because of the simplicity of his probabilistic model, he makes a number of implicit assumptions which are far from justified.

One of the assumptions is that if the definitions of the relevance classes are changed (say from two to three such classes), this will have no effect on the expected or mean relevance (on his relevance scale) of a given document; this in spite of the fact that *the form of his relevance scale* is highly dependent on the number of relevance classes. For example, suppose that the two relevance scales under consideration are:

(a) relevant/non-relevant
(b) highly relevant/partially relevant/non-relevant

Gebhardt's scale assumes that these categories have quantitative values assigned to them, within (and implicitly spanning) the range [0,1]. So the relevance scales become:

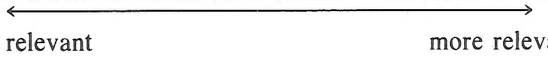


†Saracevic has recently updated his review[5].

It is clear that, if the "partially relevant" category is used at all, the expected relevance of a given document must be lower on scale (b) than on scale (a); thus his assumption is not justified.

In order to solve this problem, we must think a bit more deeply about the process of judging relevance, and how a judge might take account of the defined categories. If we are to make comparisons between scales (as Gebhardt tries to do), then we need a model which specifies some properties of the process that are invariant under such a change. So we must assume that there is an underlying property of relevance, independent of the categories, and that the categorization is a subsequent (conceptually separate) process. In order to allow for any possible number of categories, we must assume that the underlying property is continuous. Such a model has been proposed elsewhere[6]: the continuous underlying property is there called "synthema".

We assume, then, that the relevance of a document to a question is a continuous variable:



and that the experimenter, by specifying certain categories, delimits portions of the scale: e.g.

non-relevant | partially relevant | highly relevant

We can now consider the relationship between the present model and that recently proposed by COOK[7]. Cook is concerned with modelling the relevance judgement process in somewhat more psychological detail; in his model the continuous underlying variable appears as an explicit measure of the "total value" of a document or message. Cook considers only dichotomous judgements: thus the categorization of the scale is achieved by defining a single threshold between the relevant and the non-relevant. Cook considers this threshold to be a fundamental characteristic of the individual who asks the question. This seems a somewhat restricted view, since it is certainly possible to get judges (including the original questionner) to use more than two categories: thus the individual's definition of a boundary on the scale is at least to some extent under his conscious control. However, Cook's idea of an explicit representation of the underlying variable, although not pursued further in the present paper, is a valuable one.

PROBABILITY IN THE MODEL

So far, the model is deterministic and contains no probabilistic ideas. (We need probabilistic ideas, as Gebhardt rightly says, in order to understand inconsistencies between judges.) How should we introduce probabilistic ideas into this model?

The central thesis of this paper is that we need to recognize two different types of "error" or probabilistic variation. The first type concerns the definition of the categories (or rather of their end-points): however carefully the experimenter tries to define what he means by "relevance" (i.e. what sort of answer to the question should be accepted), there is bound to be some ambiguity in the interpretation placed on this definition by a judge, in the context of his particular question. The second type of error concerns the placing of a document on the continuous scale (i.e. the actual judging of a specific document). The final judgement given by the judge, that this document falls in this category, is presumably the consequence of both these processes combined.

The final model proposed, then, is as follows:

(a) There exists, in the juror's understanding of the problem or request, a continuous underlying variable of relevance or synthema.

(b) The experimenter, in defining the relevance categories to be used, is attempting to specify intervals on this continuous underlying variable.

(c) In making specific relevance judgements (according to the specified categories), the juror is firstly assessing the position of each document on the continuous underlying scale, and secondly comparing this assessment with his understanding of the relevance intervals.

(d) Therefore: if we are to compare the relevance judgements of different jurors (to the same request), we should take account of two possible sources of variation:

Type I: Variations in the jurors' interpretations of the end-points of the relevance intervals;
Type II: Variations in the jurors' assessments of the documents on the underlying scale.

So, by considering the judgement process in very slightly more detail than Gebhardt, we are forced to the conclusion that the observed variations in relevance judgements are the result of a combination of two different sorts of variation. It seems unlikely that we will be able to make much quantitative sense of the observed variations without trying to model explicitly these two distinct sources.

SARACEVIC[4] makes the following general comment about work on relevance:

"In much previous work the concept of relevance and the concept of relevance judgement were confused with each other".

It will be clear that one of my criticisms of Gebhardt is exactly that he perpetuates this confusion. My model is a direct attempt to elucidate the situation, by specifically identifying the concept of relevance and the process of judgement as two separate components of the problem.

The discussion above recalls the early work of MARON and KUHNS[8], which has been much quoted. Maron and Kuhns 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. This definition unfortunately compounds the confusion mentioned above, between the concept of relevance and that of relevance judgement, by introducing a third component (the index terms), and by slipping in a basic undefined variable ("satisfaction") which is clearly assumed to be dichotomous. We need to elucidate the three components individually before we start trying to combine them; in terms of my model, the three components are:

(a) a basic underlying variable (assumed to be continuous);
(b) a relevance-judgement process, which is subject to the two types of error defined above;
(c) an attempt by the system to make a prediction of the relevance of a document to a question.

Probabilistic ideas have a part to play in both (b) and (c), though in rather different forms.

QUALITATIVE ANALYSIS OF THE MODEL

We now return to the model of the relevance-judgement process proposed above.

As it stands, the model is only qualitative: we have not attempted to introduce any quantitative ideas into it. However, it is possible to examine some experimental results in the light of this model in its qualitative form. For this purpose, I will make use of Saracevic's[4] excellent review of the general area.

One of Saracevic's conclusions is:

"Most significantly, although the rating of degrees of relevance may be scattered, the *relative* positions of documents as to their relevance, especially among the documents with high relevance, may be expected to be remarkably consistent even among groups of judges with differences in subject education" (my emphasis).

In terms of the model proposed above, the implication of this conclusion is clear and unambiguous: it is that the first type of error or probabilistic variation (that concerned with the definitions of the end-points of the categories) is a very much more important source of variation than the second (that concerned with the assessment of an individual document on the underlying continuous scale).

The important aspect of this interpretation lies in the fact that a Type I variation affects all documents assessed by one juror in the same way. Gebhardt's model contains no mechanism

for such an effect. I therefore take the result as experimental support for the model proposed above, at least in contrast to Gebhardt's.

Another conclusion drawn by Saracevic concerns the characteristics of the judges:

> "It ... may be expected that the greater the subject knowledge in a group, the fewer documents will be judged relevant; that is, the judgement will be most stringent; conversely the less the subject knowledge, the more lenient the judgement is".

Once again, this is easily interpretable in terms of my model: the interpretation is that the first type of variation correlates with the subject knowledge of the judge. Saracevic notes other variables which have the same effect.

There are, of course, many variables identified by Saracevic which affect the extent of agreement between judges. These can be interpreted in terms of either type of variation (or both) in my model. In some cases, a more detailed analysis of the existing experimental results may indicate which of the types a given variable affects; in others, additional experiments may be necessary.

QUANTIFICATION OF THE MODEL

As mentioned above, Gebhardt's model assumes that the categories of relevance can be assigned quantitative values in the range [0, 1]; he makes extensive use of these quantities in defining measures of performance. He observes that his model is not changed by any "continuous monotone mapping of the interval [0, 1] onto itself", although of course his measures of performance *would* change under such a transformation. He appears to regard this problem simply as one of definition: "[One question is] to find and define a standard scale for X".

It seems to me, however, that the problem is more fundamental, in that as we have seen above, it is the very definition of a scale that gives rise to some of the variations in relevance judgement. It seems that the fundamental property of relevance is best reflected in the *relative* positioning of different documents rather than their *absolute* rating. Any attempt to use quantitative values in this way is likely to fall foul of this problem, by depending on the essentially arbitrary absolute ratings as well as the more important relative positions. In particular, Gebhardt in effect *defines* the absolute relevance of a document as the mean (expected value) of the relevance assesments made by different judges; this measure clearly depends on the arbitrary nature of the scale.

In my model, the absolute relevance is taken as a basic variable, and therefore this particular problem does not arise. However, there remains the problem of making inferences about this absolute variable from observations of actual relevance judgements; for this purpose some degree of quantification is required. At present, I can only indicate what would be the necessary components of such a quantification.

We need first some idea of how the documents in a collection are distributed with respect to the underlying absolute relevance variable. (In effect; we need a generalization of the traditional "generality" measure—i.e. the proportion of documents in a collection that are relevant). Secondly, we need some idea of the error-structure for the two types of error. There is hope that we might be able to discover something about these error-structures from previous experiments on relevance, but as mentioned above there are likely to be problems isolating the two different kinds of error.

IMPLICATIONS FOR EVALUATION

Gebhardt seeks to define, on the basis of his relevance scale, some new measures of retrieval performance. In particular, he defines two measures which are generalizations of recall and precision.

Two problems arise immediately with these definitions. The first is that if we admit several grades (or a continuum) of relevance, just two parameters are not enough to describe completely the results of a test. Secondly, Gebhardt's measures are defined in terms of his specific quantitative scale of relevance, and therefore suffer from the problem mentioned above, the arbitrariness of this scale.

Elsewhere [6], I have proposed a definition of a general measure of *IR* system performance, which solves the first, and in part the second, of these two problems. The measure is:

$\theta(R)$, the probability that a document will be retrieved, in response to a particular question, given that it is of grade R relevance to that question.

If we assume that relevance is a continuous variable, then this measure is a function which describes the behaviour of the system for *any* value of this variable, thus providing a more complete description than Gebhardt's two measures.

Further, in its *definition* this measure is independent of any particular quantification of the relevance model. I stress *definition* because as soon as we want to make any inferences about the measure, such as estimating it, from the results of a test, we may well have to make use of particular quantitative models. However, I would stress the importance of making the definition independent of the quantitative model in this way, at least until we have a better basis for assuming a particular model.

It will be noticed that the probability function defined above is related, in the two-relevance-grades case, to the usual measures recall and fallout. However, it cannot be stressed too strongly that recall and fallout, defined in the usual way as proportions, are *estimates* of these probabilities. More important, they are not necessarily the best estimates (the estimation problem is considered in more detail elsewhere [6]).

In the context of the present relevance model, if we had perfect relevance information (i.e. the exact position of each document in the continuous relevance scale), then estimating the probability function $\theta(R)$ would be a form of regression problem. Thus in statistical terms, our problem becomes one of regression analysis *with error in the independent variable.* The fact that there are methods for dealing with such statistical problems encourages us to think that we might be able to deal with ours; but the peculiarly complex nature of the error-structure in our case makes it an extremely difficult statistical problem.

CONCLUSIONS

(1) Gebhardt's "simple probabilistic model" is too simple to represent adequately some characteristics of the relevance-judgement process.

(2) A more complex model suggests that there are two very different types of "error" involved in observed variations in relevance judgements.

(3) The proposed model receives some support from previous experimental studies of relevance.

(4) A much more detailed quantitative understanding of relevance, through theory- and model-building, is required. Such quantitative modelling must go hand-in-hand with experimental studies of relevance.

(5) Until we have this quantitative understanding, the full implications of relevance-judgement variations for retrieval tests are difficult to assess.

REFERENCES

[1] F. GEBHARDT, A simple probabilistic model for the relevance assessment of documents. *Inform. Proc. Man.* 1975, **11**, 59.

[2] C. A. CUADRA and R. V. KATTER, Opening the black box of "relevance". *J. Docum.* 1967, **23**, 291.

[3] A. M. REES *et al. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching.* Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, Ohio (1967).

[4] T. SARACEVIC, The concept of "relevance" in information science: a historical review. In *Introduction to Information Science* (Edited by T. SARACEVIC), pp. 111–151. Bowker, New York (1970).

[5] T. SARACEVIC, Relevance: A review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inform. Sci.* 1975, **26**, 321.

[6] S. E. ROBERTSON, *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems.* Ph.D. Thesis. University of London (1976).

[7] K. H. COOK, A threshold model of relevance decisions. *Inform. Proc. Man.* 1975, **11**, 125.

[8] M. E. MARON and J. L. KUHNS, On relevance, probabilistic indexing and information retrieval. *J. Ass. Comp. Machinery* 1960, **7**, 216.