Information Technology: Research and Development (1982), *1* (1-21)

PROBABILITY OF RELEVANCE: A UNIFICATION OF TWO COMPETING MODELS FOR DOCUMENT RETRIEVAL

S. E. ROBERTSON

Centre for Information Science, The City University, London, UK AND

M. E. MARON AND W. S. COOPER

School of Information Studies, University of California, Berkeley, CA 94720, USA

(*Received 15 August 1981*)

ABSTRACT

The basic ideas underlying two probabilistic models for document retrieval are analysed. Maron and Kuhns' model (Model 1) is seen as grouping users together in order to compute a probability of relevance for a given document; Robertson and Sparck Jones' model (Model 2) groups documents together in order to compute a probability of relevance for a given user. A unified theory is presented, which contains four specific models: Models 1 and 2, a lower level Model 0 (which groups both documents and users) and a higher level Model 3. In Model 3, the individual event of a user making a relevance judgment on a document is regarded as the interaction of two sets of events: individual user with group of documents, and group of users with individual document. Some possible solutions to Model 3 are discussed. A simplified system embodying an application of the unified theory is described. Matters raised include: the nature of indexing and query formulation; the complementarity of Models 1 and 2; the notion of probability of relevance and implications for the probability ranking principle.

1. INTRODUCTION

The central concept in any proper formulation of the document retrieval problem must be the concept of relevance. Relevance is a relationship that may (or may not) hold between a document and a person in search of information: if that person wants the document in question, then we say that the relationship of relevance holds. Whether or not a given person judges a given document as relevant is a function of a large number of variables concerning that document (e.g., what it is about, how and when it was written, etc.) and also, of course, concerning that inquiring patron (e.g., what he already knows or believes, the problem that has motivated his search for information, his level of education). Because of all the many variables involved, it is virtually impossible to make a *strict* prediction as to whether the relationship of relevance will hold between a given document and a given person. Therefore, the problem must be approached probabilistically. And in fact, two separate and quite different probabilistic models for the document retrieval problem have been described in the literature of this field.

One model was proposed by Maron and Kuhns (1960), the other by Robertson and Sparck Jones (1976). Document retrieval systems based on each of these two probabilistic models would operate as follows: given a request for information, the system computes for each document of the collection its *probability of relevance* for the inquiring patron. It then produces a ranked list (in descending order) of the corresponding documents where the ranking corresponds to the computed value of the probability of relevance for each document. By providing the inquiring patron (or some intermediary searcher) with a ranking of the documents according to their probability of relevance, the system thus provides that person with an 'optimal' strategy for searching through the output. In the first (Maron and Kuhns) model, binary subject indexing is replaced by weighting indexing. In the second (Robertson and Sparck Jones), binary query terms are replaced by weighted query terms. However, in both systems the weights are interpreted as estimates of precisely defined (but different) probabilities. Can these two different probabilistic models for the document retrieval problem be unified under a single theoretical framework?

In order to develop a statistical (probabilistic) model of retrieval that is fruitful in the sense of yielding useful rules (or principles) for retrieval, it is necessary to group the individual events (described above) into classes, so that information from some members of each class can be used to make predictions about other members. But the individuality of the event lies in the individuality of its two components, namely the individual properties of the document in question and the individual properties of the person in search of information. Thus one may, as a first step, continue to regard the document as an individual but group together people according to the properties of their information needs. On the other hand, one may group documents together (according to their properties) and regard the person (in search of information) as an individual. And, in fact, these two different approaches to the problem are exactly the ones taken by the two probabilistic retrieval models described above.

The purpose of this paper is to present a unified theory which combines both of the above approaches. We shall show that in addition to the two existing probabilistic models two further models can be identified. A low-level model groups both documents and people. (It appears that such an approach is already implicit in some uses of the two earlier models.) The high-level model regards the individual event as the interaction of two groups of events, namely, the individual person taken together with the group of similar documents, and the individual document taken together with the group of similar people (i.e., people who have similar information needs). This paper will examine the nature of the grouping process, and the statistical character of the unified formulation and the high-level model.

2. THE DOCUMENT RETRIEVAL PROBLEM

2.1 A first formulation of the problem

Here is a preliminary way to think about the document retrieval problem. There exists a corpus (or collection) of writings (or documents). These documents can vary greatly in size, style of writing, depth and completeness of coverage, quality, subject content, date of publication, affiliation of author, etc. However, each can be said to 'contain' information of some sort. In addition there exists a population of people called 'patrons' who periodically seek recorded information for a variety of purposes. For any particular patron (at any particular time) the information that he wants or needs may be contained in one or more of the documents of the collection. How can each patron find the information that he wants? Or, stated differently, how can each patron find all and only those documents that contain the information that he wants when he wants it? This, in a nutshell, is the document retrieval problem.

2.2 The meaning of relevance

We have said that the function of a document retrieval system is to retrieve all and only those documents that the inquiring patron wants (or would want). We have assumed implicitly that relative to every search for information each patron either wants a given document or else he does not want it. We also have assumed that any document that he wants, he continues to want even after seeing other documents. We now make these assumptions very explicit. If a document is one that the patron wants, then we call it a 'relevant' document, relative to his (i.e., the patron's) search for information. If it is one that he does *not* want, then we call it a 'non-relevant' (or 'irrelevant') document. The relevance of one document does not depend on which other documents he has seen. The use of the terms 'relevant' and 'non-relevant' enables us simply to talk about the document retrieval problem in a more economical fashion. It now enables us to formulate the document retrieval problem in terms of the following question: how can one design a document retrieval system which can accept requests for information (from an inquiring patron) and then divide the collection into two exclusive and exhaustive sets, namely, into the set of relevant documents which it selects and retrieves and the set of non-relevant documents which are not retrieved?

For years the concept of relevance has been the subject of much discussion and controversy in the literature of this field. At times it has been defined as a relationship between a document and a topic. It has also been defined as a relationship between a document and a search query. And there have been other definitions. However, we feel that the concept of relevance, in its document retrieval sense, is best defined, as described above, as a relation between a document and a person, relative to a given search for information. Given this interpretation, the set of relevant documents is simply the set of those documents that a patron wants.

2.3 Enter probability of relevance

Now let us look more closely at this central relationship of relevance that may or may not hold between a document and a person in search of information. More specifically, we now must ask how to decide, given a document and a person (with a need for information), whether that document would be judged relevant by that person. What are the functions that influence whether or not the relationship of relevance holds between a given document and a given person?

A document is a complex entity. First of all it has a content, *viz.*, what it is about, what it says, how what it says is written, whether the information is well-organized, clear, consistent, up-to-date, etc. In addition and separate from the content is the context of a document, *viz.*, its author and his background and affiliation, where the document was published and when, what it cites, etc. All of the above factors (and many more) may influence whether a person in search of information will judge *that* document as relevant.

And what about the person doing the judging (deciding) about relevance. From an information point of view he is an even more complex entity than the document. He is looking for information for some definite or perhaps vague reason. He may or may not know clearly what will satisfy him. He may not know how to describe what information he wants. Furthermore, he comes to the search situation with particular knowledge, *viz.*, with the content of his own internal memory. We speak of his internal cognitive map as representing what he knows or believes, and of the gaps in that map as representing areas of his ignorance. The terrain represented by his cognitive map, his problems and objectives, and also the very nature of his cognitive system (his intelligence, style, etc.) will influence whether he will judge a given document as relevant.

Thus we see that there are dozens and dozens of extremely complex (and poorly understood) factors (or properties) of a document and of a person that together determine whether that person would want that document. It is for this reason that relevance cannot be predicted with any certainty. Thus no document retrieval system designed to 'retrieve all and only the relevant documents' can succeed completely. And, in fact, any system designer who poses the document retrieval problem in the above terms (*viz.*, to retrieve all and only the relevant documents) has formulated his problem improperly.

Given what we have just said about the complex set of factors which influence relevance, we see that the document retrieval problem must be framed as a problem of evidence and prediction. That is to say, the problem is to combine our knowledge (or the system's 'knowledge') of the various properties of a document with that of the properties of a person in search of information and from that combination of evidence attempt to predict whether or not the relationship of relevance holds between the document and the patron. In other words (given our assumption about relevance), the document retrieval system must combine available evidence in order to compute (or assign a value to) the probability of relevance for any given document and patron. If the system can compute these probabilities, then the values can be used to rank the documents in question, i.e., it can provide the inquiring patron with ranked output, where the documents are ranked in descending order by their probability of relevance.

2.4 How to interpret probability of relevance

Because relevance is a relationship between a document and a person in search of information, it has two aspects (or poles): the document aspect and the patron aspect. These two poles of the relevance relationship point to two different ways of making predictions about relevance. In order to predict relevance one must be able to do one (or both) of the following:

1. Correlate documents with (the information properties of) those people who would judge them relevant.

2. Correlate people with (the information properties of) those documents that they would judge relevant.

These two approaches are represented by the two different probability models that have been discussed in the literature of this field.

In 1960 Maron and Kuhns proposed a probabilistic model (which we will refer to as Model 1) for the document retrieval problem. Model 1 interprets the situation as follows: a patron submits a query (call it Q) consisting of some specification of his information need. Different patrons submitting the same stated query may differ as to whether or not they judge a specific document relevant. The function of the retrieval system is to compute for each individual document the *probability* that *it* will be judged relevant by a patron who has submitted a query Q. Thus, in Model 1, probability of relevance is computed relative to evidence consisting of the type of query that the patron has submitted.

Maron and Kuhns argued that probability in their model be interpreted in its frequency sense. This means, for example, that in the case where the query consists of a single term (call it I_j^Q), the probability that a given document d_m will be judged relevant by the patron submitting I_j^Q is simply the ratio of the number of patrons who submit I_j^Q as their search query and judge d_m relevant, to the number of patrons who submit I_j^Q as their search query.

Thus in this model, documents and queries are looked at 'historically' over time, i.e., probability of relevance for each document is interpreted as the number of times that it is judged relevant by a patron whose search term is I_j^Q, divided by the number of times that a patron submits I_j^Q as a search term.

Now consider a different interpretation (called here Model 2) of the document retrieval problem. Model 2 is based on the work of Robertson and Sparck Jones (1976), and interprets the document retrieval situation as follows: documents have many different properties; some documents have all the properties that the patron asked for, and other documents have only some or none of those properties. If the inquiring patron were to examine all the documents in the collection he might find that some having all the sought after properties were relevant, but others (with the same properties) were not relevant. And conversely, he might find that some of those documents having none (or only a few) of the sought after properties were relevant, others not. The function of the document retrieval system is to compute the probability that a document is relevant, given that it has one (or a set) of specified properties. Thus in Model 2 (as opposed to Model 1) probability of relevance is computed relative to evidence consisting of the properties of documents.

Model 2 (like Model 1) also adopts a frequency interpretation of probability. This means, for example, that the probability that a document having property I_j is relevant is the ratio of the number of relevant documents having the property I_j to the total number of documents having the property I_j.

Here are two different models with different ways of interpreting and computing probability of relevance. In Model 1, probability of relevance is computed relative to the set of queries. In Model 2, probability of relevance is computed relative to the set of document properties. Because probability is a relationship relative to evidence, two probabilities of the same hypothesis relative to different evidence could have different values. Again, in Model 1, the evidence consists in the search (query) term (or terms) submitted by the patron. In Model 2, the evidence consists of the properties of (i.e., index terms assigned to) the document.

Vastly different consequences flow from systems based on these two different models. The indexing and query formulation procedures would be different; the output rankings would be different; and we could expect (and would find) that the overall retrieval effectiveness of two such systems would differ.

Let us now look more carefully and in more detail at each of these two models and at their implications and consequences. Only then will we ask whether and how these two probabilistic models for the document retrieval problem might by synthesized to form a single unified theory.

3. TWO PROBABILISTIC MODELS FOR THE DOCUMENT RETRIEVAL PROBLEM

3.1 Model 1: the formal presentation

Model 1 was first proposed by Maron and Kuhns (1960) and extended by Cooper and Maron (1978). The present description is close to that given in these papers, though with some changes of notation.

We assume that at any time there exists a collection of n different documents which we designate $d_1, d_2, d_3, \ldots, d_n$. We assume also that there exists a vocabulary of s different terms which we designate $I_1, I_2, I_3, \ldots, I_s$. For simplicity and clarity of exposition we will consider here the simplest case, namely, where the query consists of a single term I_j. Given a patron's input query, the function of the system is to compute the value of $P(d_m^R|A,I_j^Q$ for each document d_m. $P(d_m^R|A,I_j^Q)$ is the probability that if a patron submits I_j as his query, then he will judge document d_m as relevant.

The classes A, d_m^R and I_j^Q that appear in the probability expression $P(d_m^R|A,I_j^Q)$ are defined as follows:

A = the class of uses of the system by a patron in search of information.
I_j^Q = the class of single term queries of the type I_j.
d_m^R = the class of events each of which consists of a patron judging the mth document d_m as relevant.

Given its frequency interpretation, the value of the above probability is equal to the number of elements in the class 'A,d_m^R,I_j^Q' divided by the number in the class 'A,I_j^Q', i.e.,

$$P(d_m^R|A,I_j^Q) = \frac{N(A,d_m^R,I_j^Q)}{N(A,I_j^Q)},$$

where '$N(\quad)$' stands for 'number of' and ',' denotes logical intersection.

The theorem of Bayes allows us to write the following equality:

$$P(d_m^R|A,I_j^Q) = \frac{P(d_m^R|A)\cdot P(I_j^Q|A,d_m^R)}{P(I_j^Q|A)}. \tag{1}$$

For any given query I_j^Q, the denominator of (1) is a constant and thus we may rewrite (1) as follows:

$$P(d_m^R|A,I_j^Q) = k.P(d_m^R|A)\cdot P(I_j^Q|A,d_m^R).$$

$P(d_m^R)$ is the probability that the mth document d_m will be judged relevant, independent of the query term that the patron submits. (Sometimes it is called the *a priori* probability of relevance.)

$P(I_j^Q|A,d_m^R)$ is the probability that if d_m were to be judged relevant by a patron, then he would be using I_j as his query term.

Given the values of $P(d_m^R|A)$ and $P(I_j^Q|A,d_m^R)$, the system can merely take their product and use that value (after it has been normalized) to rank the output documents. Where would the values of those probabilities come from?

Feedback from users could provide statistics needed to estimate the values of $P(d_m^R|A)$. Similarly, user feedback could provide statistics needed to estimate the values of $P(I_j^Q|A,d_m^R)$. However, in lieu of actual past statistics, an indexer could attempt to estimate the values of $P(d_m^R|A)$ and $P(I_j^Q|A,d_m^R)$. In fact, according to this model, called by Maron and Kuhns 'Probabilistic Indexing', the task of the indexer is precisely defined; his job is to estimate the values of $P(I_j^Q|A,d_m^R)$, for all I_j and all d_m and then to assign those index terms I_j to the corresponding documents with the values of those estimates. Thus we end up with a theory of weighted indexing, where the weights are estimates of the probability $P(I_j^Q|A,d_m^R)$.

3.2 Model 2: the formal presentation

We now consider the second probability model, called here Model 2. The origins of this probabilistic model rest on the work of Robertson and Sparck Jones (1976). Its most complete formulation was presented by van Rijsbergen (1979). In the description that now follows we have kept the key ideas of their model; however, our formulation is somewhat different. What is important for our purpose here is not the computational details of how to calculate probability of relevance, but rather the way probability of relevance is interpreted.

Let U be the class of events each of which consists of a document being judged relevant by the same patron, relative to his information 'need'. Then not-U, designated '$\bar{U}$', is the complement of U and consists of the class of events each of which consists of a document being judged not relevant by that same patron.

Every document has one or more of s possible properties, which we designate I_1, I_2, I_3, . . ., I_s. These properties are often the properties of being assigned some term by an indexer. The problem of the document retrieval system, under this model, is to compute the probability $P(U|I_p,,,I_t)$ that a randomly selected document will be judged relevant, given that it possesses the properties $I_p,,,I_t$, where ',' denotes logical conjunction.

For simplicity and clarity of presentation consider a document with only two properties, I_j and I_k. In this case the problem for the document retrieval system is to compute the value of

$$P(U|I_j,I_k) \tag{2}$$

i.e., the probability that a randomly selected document which has the property I_j and also the property I_k will be judged relevant by the inquiring patron.

Probability of relevance in Model 2 may be computed as follows:

$$P(U|I_j,I_k) = \frac{P(U)\cdot P(I_j,I_k|U)}{P(I_j,I_k)}. \tag{3}$$

If one makes the simplifying assumption that the properties I_j and I_k are independent of one another relative to U, then we may rewrite (3) as follows:

$$P(U|I_j,I_k) = \frac{P(U) \cdot P(I_j|U) \cdot P(I_k|U)}{P(I_j,I_k)}. \tag{4}$$

(Expansions other than (4), based on other independence assumptions, are possible.) Let us look at what the probabilities on the right hand side of (4) mean.

We have already said that $P(U)$ is simply the probability that a randomly selected document will be judged relevant by *this* inquiring patron. We can imagine that the patron himself might make some estimate of $P(U)$ and input that estimate to the system. (As a matter of fact, the use of the model for ranking documents does not require a value to be assigned to $P(U)$.) $P(I_j|U)$ is the probability that if this patron were to judge a randomly selected document as being relevant, then that document would have the property I_j, and similarly for I_k. In order for the system to compute probability of relevance, it would need some estimate of these probabilities for each term in the query. We can distinguish at least three different ways of obtaining estimates of $P(I_j|U)$. One way would be to ask the patron himself to estimate the value of $P(I_j|U)$ or some related quantity. A second would be to establish, from the results of previous searches, term properties that could be used (by human or machine) as predictors of $P(I_j|U)$: an example is term frequency. A third way would be for the system to generate some 'trial retrieval' and then ask the inquiring patron to examine those output documents and then divide them into two classes — those that he judges relevant and those that he judges as not relevant. Given these data, fed back from the patron, the system would compute the values of $P(I_j|U)$, $P(I_k|U)$, etc., assuming of course that the sample generated by trial retrieval is adequate in size, representation, etc.

What about the third probability — the one that appears as the denominator of (4)? $P(I_j,I_k)$ is simply the relative frequency among all the documents of those that possess both properties I_j and I_k. The values for this probability are 'known' by the system, i.e., it can examine all of the records and compute the value of $P(I_j,I_k)$.

Thus we see that given the values of the above probabilities, the system could compute the probability that a randomly selected document which has properties I_j and I_k would be relevant. And, of course, it could compute the probabilities for the other three cases, namely where a document has either one or else the other of those properties, or neither.

If values of $P(I_j|U)$ are available for all possible I_j, and again given a sufficiently general independence assumption, then it is possible to compute the probability of relevance in the case where a randomly selected document has *any* number of properties.

3.3 A unifying notation

A unified model must deal with *groups* of documents and *groups* of uses. (The word 'use' is used here as shorthand for the event consisting of a person with a need for information using the document retrieval system.) Thus the event-space (needed for a unified model) must consist of all possible document–use pairs. The relevance relationship will hold between some subset of the above. Thus the notation that we need to describe this model requires the following symbols:

$A =$ the class of all (past and future) uses of the system;
$C =$ the class of all (present and future) documents in the system;
$b_k =$ an individual use;
$d_m =$ an individual document.

Thus the product set $A \times C$, which consists of the class of all use–document pairs (b_k,d_m), is the event-space. The relevance relationship holds between some of those pairs, therefore relevance is a subset of the event-space. We denote this as follows:

$$R \subseteq \mathrm{A} \times C,$$

where $(b_k,d_m) \varepsilon R$ if d_m would be judged relevant to b_k.

Since Model 1 proceeds by grouping similar uses according to the formal query presented, and since Model 2 proceeds by grouping similar documents according to the index terms assigned to each, we require a notation for such classes:

$B \subseteq A =$ class of similar uses;
$D \subseteq C =$ class of similar documents.

Without loss of generality, we can assume that

$b_k \ \varepsilon \ B$ for $1 \leq k \leq K$;
$d_m \ \varepsilon \ D$ for $1 \leq m \leq M$.

This completes the essential notation for the unified model. Notice that the *probability of relevance*, which is the central concern of Model 1, becomes

$$P(R \mid B,d_m).$$

And notice further that using this notation the Model 2 probability of relevance becomes

$$P(R \mid b_k,D).$$

In what follows, we shall define a new model, to be called Model 3. The aim of Model 3 will be to evaluate

$$P(R \mid b_k,d_m),$$

that is, to evaluate the probability that the individual document d_m would be judged relevant for the individual use b_k.

4. THE UNIFIED MODEL

4.1 Properties and predictions

In order to unify Models 1 and 2, we have first to ensure that the various entities which figure in the two models are compatible, and to understand how they relate to each other. Each model includes documents, uses, index (query) terms, an indexing operation and a query formulation operation. However, in respect of the last three entities, the two models differ substantially.

The query formulation process assumed in Model 1 might be described as 'naive'. That is, the patron is assumed to express his need for information in terms of the words available, without attempting in any way to prejudge the retrieval consequences (e.g., to guess how a suitable document might be indexed). This query formulation process takes place outside the framework of the system. Thus, as far as the system is concerned, the query terms used simply describe (are properties of) the need.

The process of indexing, on the other hand, is integral to the system. The indexer has to estimate a set of probabilities concerning not only the document itself but the various patrons who may or may not judge the document as relevant. The patrons are identified by the properties of their needs (i.e., query terms). Thus both documents and needs are represented by *need* properties.

In Model 2, the situation is reversed. Indexing takes place outside the system and involves simply describing document properties; query formulation is integral to the system and involves relating the need to document properties.

We now see that the phrases 'indexing' and 'query formulation' are ambiguous: indexing means either simply identifying document properties (Model 2), or making predictions about need properties (Model 1), and similarly for query formulation. For a unified model, we need to allow for all four distinct processes to take place. More specifically, we assume that documents and needs *have* properties which are identifiable and nameable, and that the process of identifying and naming them takes place outside the framework of the probabilistic mechanism. Within the mechanism both kinds of prediction take place. Document and need properties need not be expressed in terms of the same vocabulary; clearly, linguistic labels of some sort are appropriate in both cases, but there may be other kinds of label which will identify other kinds of property (e.g., author or cited document as property of a given document). Because of the ambiguity of 'indexing' and 'query formulation', we shall eschew these phrases, and refer to document or need properties on the one hand, and a prediction process on the other.*

4.2 Event-space and probability measure

The event-space for the unified model has already been identified as $A \times C$, i.e., the set of all use–document pairs. Unless otherwise specified, all probabilities will be assumed to be conditional on this event-space. Thus

$$P(X) \equiv P((b_k,d_m) \,\varepsilon\, X \,|\, b_k \,\varepsilon\, A,\ d_m \,\varepsilon\, C);$$

$$P(X|Y) \equiv P((b_k,d_m) \,\varepsilon\, X \,|\, (b_k,d_m) \,\varepsilon\, Y,\ b_k \,\varepsilon\, A,\ d_m \,\varepsilon\, C).$$

In cases where X explicitly consists of use–document pairs, the interpretation of

* It has been suggested (Robertson, 1977a) that Models 1 and 2 are fundamentally incompatible. The argument was based on the assumption that only one set of labels is involved, and that 'indexing' therefore has only one interpretation (as has 'query formulation'). The analysis here gets round that objection. With the present perspective, we see that Model 1 on its own simply ignores document properties; Model 2 alone ignores need properties.

these expressions is clear. However, in some cases, X may refer just to uses or just to documents. For example, we may want to consider $P(B)$, where B is a class of uses, as defined above. This will be interpreted as 'probability of selecting a pair (b_k,d_m) such that $b_k \varepsilon B$', with no condition on d_m. More formally, if $B \subseteq A$, then

$$P(B) \equiv P((b_k,d_m) \text{ is such that } b_k \varepsilon B, d_m \varepsilon C | b_k \varepsilon A, d_m \varepsilon C). \qquad (5)$$

The same holds for $D \subseteq C$. This implies that $P(A) = P(C) = 1$.

We must now specify the probability measure on this event-space. We shall assume that the probability measure on $A \times C$ is uniform; i.e., that $P(b_k,d_m)$ is independent of k and independent of m.

One major reason for making the above assumption is that we wish the unified model to reduce to either Model 1 or Model 2 under appropriate conditions. (Both models assume uniform probability measures on their respective event-spaces.) Another reason for making the above assumption is that we wish to use simple frequency data to estimate probabilities where appropriate. (Nevertheless, the probability-measure problem is not trivial, and it will be discussed further below.)

The assumption of a uniform probability measure on $A \times C$ actually has strong consequences, simply because $A \times C$ is a product set. In particular, we can deduce the following independence property:

$$P(b_k,d_m) = P(b_k) \cdot P(d_m). \qquad (6)$$

More generally, if $B \subseteq A$ and $D \subseteq C$, then

$$P(B,D) = P(B) \cdot P(D).$$

It follows that

$$P(B) = P(B|D) = P(B|d_m).$$

It is also the same as the probability of B in an event-space consisting only of A (with a uniform probability measure). Thus we can talk about $P(B)$ without concern over whether the event-space is A or $A \times C$. This fact is a justification for the (apparently) arbitrary definition (5) of $P(B)$. Similar remarks apply also to $D \subseteq C$.

4.3 Outline of the unified model

We will now consider, in more detail, a subset of the event-space $A \times C$, namely, $B \ x \ D$. B was defined as a class of 'similar' uses, and subsequently, we assumed specifically that properties of the need or use were available to the system. Therefore we adopt the following definitions:

B = the class of uses (needs) which have identical properties;
e.g., they use the same query term (or terms).
D = the class of documents which have identical properties;
e.g., the same index term (or terms) is assigned to each.

Now consider (see below) a $B \times D$ matrix; i.e., whose columns refer to individual uses b_k ε B ($1 \leq k \leq K$) and whose rows refer to the individual documents d_m ε D ($1 \leq m \leq M$). The entries (individual cells) in this matrix would contain either 1s or 0s depending on whether the document d_m would be judged relevant in use b_k or not. Thus the document retrieval problem is to make a prediction about relevance, i.e., assign a value to the probability of a 1 in a cell when the relevance judgement has not yet been made.

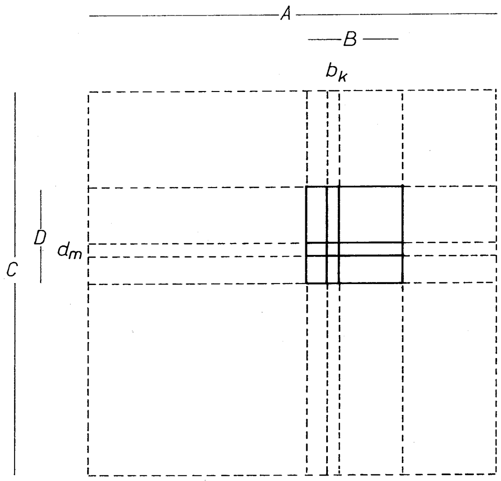

FIG 1. The $B \times D$ matrix in the context of the entire event-space

There are three kinds of information (which the retrieval system may or may not possess) and which may have a direct bearing on the above prediction of relevance. There may be (1) data describing the relations between other uses in the same class and this individual document. (This data is, essentially, the marginal information concerning the *row* of the matrix in which the cell appears.) Similarly (2) there may be data describing the relations between this individual use and other documents in the same class, i.e., marginal information about the column. Also (3) we may have data describing the relations between other uses and other documents (in the same classes). This latter is undifferentiated information about the entire $B \times D$ matrix.

If we have only the first kind of information (or perhaps the first kind together with the third kind), then we have exactly the situation appropriate for Model 1. If, on the other hand, we have only the second kind of information (or perhaps the second kind with the third), then we have the situation appropriate for Model 2. But what about the case where we have (a) only the third kind of data or (b) all three kinds?

Although it has not (to our knowledge) been formally specified before, case (a), in fact, seems to underlie some uses of earlier models. The most obvious specification of the model is: rank the documents in order of their probability of relevance, where probability of relevance is defined as $P(R|B,D)$. We shall refer to this as Model 0.

The use of frequency information to predict the Model 2 weight is an example of an implicit use of (something like) Model 0. That is, the simple application of Model 2 requires relevance feedback from the user; however, prior to obtaining such feedback we may be able to estimate the Model 2 parameters using other information such as term frequencies (see Croft and Harper, 1979). Such estimates obviously are not use-specific, but rather they depend on the fact that for different uses using the same terms there is some degree of predictability in the associated relevance judgements.

The most interesting and also the most difficult problem arises when we assume case (b), i.e., when all three kinds of information (discussed above) are available. What we want is a model (called Model 3) which combines all of the above information in order to compute probability of relevance. The path toward such a model is not immediately clear. In terms of the $B \times D$ matrix we want to estimate the value of a given (single) cell, given the two kinds of marginal information together with information for the entire matrix. That is to say, we have estimates of $P(R|B,d_m)$, $P(R|b_k,D)$ and also $P(R|B,D)$. We now ask: what would be a good estimate for $P(R|b_k,d_m)$?

4.4 Exact solutions to Model 3

We now assume that the three kinds of information (discussed above) are available and furthermore we assume that there is sufficient frequency information to provide rather precise estimates for those three probabilities. Although we will not attempt to provide an estimation theory for our Model 3, we will indicate, qualitatively, some desirable properties of such a theory.

Because we have only marginal data, we require what might be called a 'non-interaction' model. An obvious kind of non-interaction model is that used in the χ^2 test. It would take the following form:

$$P(R|b_k,d_m) = \frac{P(R|B,d_m) \cdot P(R|b_k,D)}{P(R|B,D)}. \tag{7}$$

Unfortunately, however, it is easy to construct an example for which formula (7) gives a result outside the range [0,1], i.e., it gives a value which cannot be interpreted as a probability. Before explaining this discrepancy, we shall introduce a second solution. We shall use the notation '*O*( )' to denote odds; i.e.

$$O(X) = \frac{P(X)}{1-P(X)}.$$

Using the odds notation the following solution suggests itself by analogy with (7).

$$O(R|b_k,d_m) = \frac{O(R|B,d_m) \cdot O(R|b_k,D)}{O(R|B,D)}. \tag{8}$$

We shall presently show how (8) may be derived formally. Obviously, it could be expressed in terms of probabilities rather than odds, but then its behaviour would not be so transparent. Formula (8) avoids the problem of (7) because it always will give a value in the correct range. However, it introduces a different problem, namely, that the individual cell values derived by means of (8) (for an entire row or column) do not square exactly with the marginal data on the corresponding row or column.

Why do these problems arise? Essentially because any non-interaction model must use specific independence assumptions. Some candidate assumptions would be:

$$A_1 : P(b_k,d_m) = P(b_k)\cdot P(d_m).$$

(Note that A_1 is exactly equation (6) above, and derives from fundamental assumptions about the event-space.)

$$A_2 : P(b_k,d_m|R) = P(b_k|R)\cdot P(d_m|R)$$
$$A_3 : P(b_k,d_m|\bar{R}) = P(b_k|\bar{R})\cdot P(d_m|\bar{R}).$$

The problem however is that A_1 is not compatible with either A_2 or A_3. Formula (7) is based on A_1 and A_2, and therefore it contains internal contradictions: hence the 'difficult' values. Formula (8) is based on A_2 and A_3, which are consistent between themselves, and therefore it is internally consistent.* However, because A_2 and A_3 are incompatible with A_1, this 'causes' (8) not to fit the marginal data.

What other solutions are possible? We first reformulate (8) as follows: we use '$\lambda(\ \ )$' to denote log-odds, i.e.,

$$\lambda(X) = \log\ O(X) = \log\ \frac{P(X)}{1\text{-}P(X)}.$$

Then (8) becomes

$$\lambda(R|b_k,d_m) = \lambda(R|B,d_m) + \lambda(R|b_k,D) - \lambda(R|B,D). \qquad (9)$$

* The derivation of (8) from A_2 and A_3 is as follows. Assume all probabilities (and odds) are conditional on B and D. Then

$$\begin{aligned} O(R|b_k,d_m) &= \frac{P(R|b_k,d_m)}{P(\bar{R}|b_k,d_m)} \\ &= \frac{P(b_k,d_m|R)\cdot P(R)}{P(b_k,d_m|\bar{R})\cdot P(\bar{R})} \\ &= \frac{P(b_k|R)\cdot P(d_m|R)\cdot P(R)}{P(b_k|\bar{R})\cdot P(d_m|\bar{R})\cdot P(\bar{R})} \\ &= \frac{P(R|b_k)\cdot P(R|d_m)\cdot P(\bar{R})}{P(\bar{R}|b_k)\cdot P(\bar{R}|d_m)\cdot P(R)} \\ &= \frac{O(R|b_k)\cdot O(R|d_m)}{O(R)}. \end{aligned}$$

Inserting B,D into the appropriate places, yields formula (8).

By analogy with (9), we are able to suggest a log-linear model

$$\lambda(R\,|b_k,d_m) = \alpha_m + \beta_k + \gamma. \tag{10}$$

It is possible, in principle, to find values α_m,β_k,γ such that the individual cell estimates for a column or row match the marginal data exactly. Also (10) will always give a value in the correct range. In fact, (10) is a non-interaction version of a general log-linear model,

$$\lambda(R\,|b_k,d_m) = \alpha_m + \beta_k + \gamma + \delta_{m,k}$$

where 'non-interaction' means using the assumptions

$$\delta_{m,k} = 0 \qquad 1 \leq m \leq M \qquad 1 \leq k \leq K$$

which are a form of independence assumption. Models of this kind are proposed as a general approach to binary data by Cox (1970), chiefly on the grounds that they constitute the best and simplest equivalent for binary variables to normal-theory linear models for continuous variables.

4.5 The Maximum Entropy Principle and some further thoughts

One further solution elaborated on elsewhere (Cooper and Huizinga, 1981) is the use of the Maximum Entropy Principle. This solution seeks values of the probabilities $P(b_k,d_m,R)$ and $P(b_k,d_m,\bar{R})$ such that the entropy,

$$\sum_{\substack{1 \leq k \leq K \\ 1 \leq m \leq M}} [P(b_k,d_m,R) \log P\,(b_k,d_m,R) + P(b_k,d_m,\bar{R}) \,\log P\,(b_k,d_m,\bar{R})]. \tag{11}$$

is maximized, subject to the marginal data. The maximum entropy principle has been defended as providing the 'least prejudiced' distribution which can be specified under such circumstances (Tribus, 1969).

It turns out that the maximum entropy solution and the log-linear solution are identical. Thus we have strong reasons for adopting such a solution. Unfortunately, however, it does not appear possible to put this solution in a computationally simple analytical form. Some method of approximation is therefore called for. Iterative methods for obtaining the maximum entropy distribution are described by Gokhale and Kullback (1978); to carry out two or three steps of such an iteration would be one way of obtaining such an approximation. Another possibility is simply to use the odds formula (8) as an approximation.

It is not known, in general, how good an approximation the odds formula will give. Since the odds formula can be expressed in exactly the same form as the log-linear model, but differs from it in that the marginal totals are not correct, it is appropriate (as a first step) to establish just how far out the marginal totals are. It can be shown that how far out they are depends on the extent of variation between individual rows or columns, and on the overall probability of relevance $P(R|B,D)$; if the extent of variation is small, and $P(R|B,D)$ is neither too large nor too small, then the discrepancy is small. However, the question of whether simple, good approximations can be found requires further investigation. In the interim, we make use of the odds formula solely in order to illustrate the properties of Model 3.

4.6 Probability measure revisited

We have seen that the best solution for Model 3 is not easy to express in a simple form. Further, a candidate simple solution, the odds formula (8), depends on the independence assumptions A_2 and A_3. And these are not compatible with the assumption A_1, which itself is a direct consequence of our choice of probability measure on the event-space $A \times C$. Thus the question arises, could we choose a different probability measure, with which A_2 and A_3 were compatible? If so, then the odds formula would be identical with the two 'good' solutions (10) and (11).

Abandoning the uniform probability measure would mean abandoning Model 1 and Model 2 in their present form, and also abandoning simple frequency data. Although we cannot see any logical argument against such a course of action, it does seem likely to be counter-productive. Therefore, we do not pursue that course of action in this paper, but will continue to regard the odds formula as a useful (but not necessarily a highly accurate) approximation to a better solution.

4.7 Generalization to utility theory

Although the details will not be explored here, it is a straightforward task to generalize the unified model to take account of degrees of relevance, or document 'utilities'. The output ranking of documents is thus in descending order of estimated expected utility rather than estimated probability of relevance. For a utility-theoretic development of Model 1 see Cooper and Maron (1978).

5. A SIMPLE APPLICATION OF THE UNIFIED MODEL

5.1 First remarks

In this section we shall describe a retrieval system designed on the basis of the unified probabilistic model presented above. This is not intended as a proposal for a real system. Rather, it is deliberately simplified in a number of respects for the following reasons: firstly, we want to bring out what we consider the essential features of the new model, and secondly, some development work would be required before a realistic retrieval system using these ideas could be designed. In the course of the discussion that follows we shall indicate the simplifications built into this example, and some aspects of the model which require further development.

5.2 The basic system

We assume that the properties of the documents and needs have been identified and named, as discussed above. Furthermore, we assume that the named properties take the form of one word per document or need. (Needless to say, this is one of the unrealistic simplifying assumptions referred to above.) Thus the documents (and also the uses) are classified into a number of exclusive classes according to the single word used. The vocabulary of document property names need not be the same as that of need property names; however, having the same vocabulary for both may be useful, as we shall presently see.

Assume further that the retrieval system has been operational for some time and that relevance feedback has been obtained from users. Thus we assume that for each class of uses B, and for each class of documents D, there exist frequency data from which accurate estimates of $P(R|B,D)$ can be made. Furthermore, assume that for each individual document d_m that has been in the system long enough, we have an estimate of

$$P(R|B,d_m).$$

Since we will be using the odds formula (8) we shall use the odds instead of the corresponding probabilities.

When a new use b_k is made of the system, an iterative search is performed; initially it is identified only as a member of a class of uses B. Each document in the system belongs to a class D; also there *may* be enough data for any document so that it can be treated as an individual d_m. Thus, for the initial search each document is given a value which is either

$$\begin{array}{rl} \text{(a)} & O(R\,|\,B,D) \\ \text{or (b)} & O(R|B,d_m), \end{array}$$

according to whether that document is a recent or else an older acquisition. The documents of the collection are then ranked by these values and those with the highest computed values are retrieved for relevance evaluation.

The second pass for b_k involves using the relevance judgements obtained from the first pass in order to estimate

$$O(R|b_k,D).$$

The values initially assigned to the documents then are corrected by a factor

$$\frac{O(R|b_k,D)}{O(R|B,D)} \tag{4}$$

and the documents are reranked.

It can be seen that for documents in class (a) above, the corrected value reduces to

$$O(R|b_k,D).$$

For those documents in class (b) above, the corrected value is the odds formula solution to Model 3. Thus we can describe the system as follows:

1. Given data about the individual document and the individual use, it is a full Model 3 system.
2. Given data about the individual document, but not the individual use, it is a Model 1 system.
3. Given data about the individual use, but not the individual document, it is a Model 2 system.
4. Given individual data about neither documents nor uses, it is a Model 0 system.

Let us now turn and consider some extensions to the simple system.

5.3 Bayesian ideas

For each of the probabilities required by the system, we either assumed that we had adequate data to estimate it accurately or else we proposed a method that did not involve estimating it at all. In general, one would be more likely to have some partial (small sample) data from which could be obtained a rather imprecise estimate. How could the system be adapted so as to deal with this aspect of the real situation?

We seek Bayesian methods which can be used to modify estimates whenever data are available, given prior expectations. For example, if we already know $P(R|B,D)$, this can serve as a prior expectation for $P(R|B,d_m)$ and it can be modified gradually as we acquire data about d_m in relation to uses B.

What prior expectation might be advanced for $P(R|B,D)$? We need this not just when the system first starts up, but for *some* pairs B,D, we will need it for a long time thereafter, because data about some pairs will be less readily forthcoming than about others. If the same vocabulary is used for documents and uses, the obvious prior expectations would be:

$$P(R|B,D) = \begin{cases} 1, & \text{if } B \text{ and } D \text{ are associated with the same term} \\ O, & \text{otherwise.} \end{cases}$$

The use of this technique would, unfortunately, have one major disadvantage, namely, for any pair (B,D) for which the prior $P(R|B,D) = O$, no use in B would ever retrieve a document in D. Hence no document in D would ever be judged for relevance against a use in B, and no data would ever be obtained which might serve to modify the prior $P(R|B,D) = O$; the situation would be self-perpetuating. However, it might be possible to find alternatives, such as techniques based on thesaural connections between terms. The problem might be alleviated in a system that allows multi-term indexing and/or searching (see below).

This problem is an extreme instance of a more general problem in this area, namely, the problem of bias in estimation. In general, in order to get unbiased estimates of any parameter, one seeks to obtain random samples of events. However, in any realistic retrieval situation, one starts with those pairs which are most likely, according to prior information, to be judged relevant. Thus all small-sample estimates are biased. This is a vexing problem for those who now are developing Model 2 retrieval systems involving user feedback, and no general solutions are yet forthcoming.

Finally we point out that, in general, Bayesian methods require not merely prior expectations (i.e., means), but prior probability distributions. (The amount of spread in the prior distribution may be regarded as a measure of the degree of confidence in the prior expectation, relative to incoming data.)

5.4 Multi-term representations

Our assumption for the basic system that only single-term descriptions of documents and needs are permitted clearly is unrealistic. In fact, Model 2 depends on the use of many terms for indexing and especially for query formulation.

As can be seen from the prior discussion, in any multi-term situation the exclusive classes which are required for these probability models consist of classes defined by combinations of terms — not by single terms alone. Thus *any* combination of terms (implying negation of terms not explicitly included) defines a unique class of items (documents, uses, or whatever), and these classes are exclusive in that no item can belong to more than one.

A problem with such classification is that each class is likely to contain only a very small number of items. (Indeed, most potential 'classes' will be empty.) Hence any collection of statistical data on this basis is likely to be fruitless: it is highly unlikely that one would be able to deduce anything about a new event (document and/or use) from previous ones.

A technique adopted by Model 2 in order to circumvent this problem may be useful in the unified model. This technique was to consider the properties of an item defined by a combination of terms (present and absent) to be derived from the properties of the terms taken singly. This may be done most simply by making strong assumptions about independence between terms; it is also possible to allow a limited form of dependence. It must be stressed, however, that trying to allow for any dependence relation would lead us back to the previous situation, of not having enough items in each class. Thus, some kind of independence assumption is central to this type of model.

Multi-term operation would help to get over the problem of finding appropriate prior expectations for $P(R|B,D)$, in that it would not necessarily matter if some of these were taken as zero, since particular documents in this class might be retrieved for particular uses in this class by virtue of the presence of other terms in both.

5.5 Humans in the basic system

The basic system as described was seen as a purely mechanical device: given the input of document and need properties and relevance judgements, the system would automatically collect the appropriate statistics and apply the appropriate formulae. We could, however, imagine human indexers and searchers playing a more direct part in the operation of the system. Thus there could, for example, be a human indexer making explicit predictions of $P(R|B,d_m)$ with Model 1. Equally, we may get searchers to make predictions of $P(R|b_k,D)$, as suggested recently for Model 2 (Cooper and Huizinga, 1981). The unified model would accept either or both kinds of information, either without frequency data or as a Bayesian prior expectation to be modified by frequency data as discussed above. Should both kinds of estimate be available, Model 3 indicates how they should be combined.

It should be noted, in the light of the earlier discussion on 'indexing', that such a system would actually include two quite different kinds of 'indexing' operations: the identification and naming of document properties, and the prediction of the properties of needs to which the document may be appropriate (i.e., the estimation of $P(R|B,d_m)$). Even though one might imagine a system in which the two processes are performed by the same person at the same time, they must be regarded not only as distinct operations, but as having quite different characters. Similar remarks apply to the two aspects of 'query formulation' discussed earlier.

6. CONCLUSIONS

6.1 Probability of relevance

The major conclusion to be derived from the analysis presented in this paper is that 'probability of relevance' has several different interpretations (or meanings), depending on the class of events to which this probability applies. In Model 1, probability of relevance $P(R|B,d_m)$ is a relationship between an individual document and a class of uses. In Model 2, probability of relevance $P(R|b_k,D)$ is a relationship between an individual use and a class of documents. In our (newly discovered) Model 0, probability of relevance $P(R|B,D)$ is a relationship between a class of uses and a class of documents. Finally there is our new Model 3 where probability of relevance $P(R|b_k,d_m)$ is a relationship between an individual document and an individual use. We can depict the relationships among these four models with Figure 2.

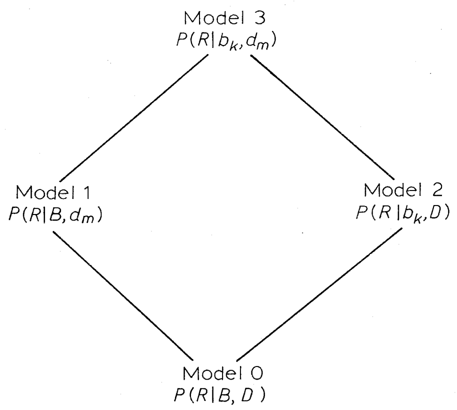

Fig. 2. Components of the unified model

The value of probability of relevance, like any probability, is relative to the evidence. In Model 1 and in Model 2, the values of probability of relevance are based on different kinds of evidence. Since Model 3 in fact makes use of the Model 1 and Model 2 interpretations of 'probability of relevance', it reveals that they are complementary, rather than competing with each other. Thus, our Unified Model, which uses both kinds of evidence, shows us that it would be mistaken to consider either Model 1 or Model 2 as 'false' relative to the other; they are, as we have said, to be seen as complementary.

6.2 The Probability Ranking Principle: a final remark

A major assumption of this paper has been that the function of the retrieval system is to estimate a probability of relevance, whereby to rank documents for retrieval for a given user. This assumption is known as the Probability Ranking Principle, which states that ranking by probability of relevance will yield optimal performance (Robertson, 1977b). There exists proof of its validity under some conditions. Although there are also counter-examples, these appear to be of little importance.

The probabilistic model presented in this paper suggests that a reconsideration of the Probability Ranking Principle is needed, since there are now four different interpretations of the phrase 'probability of relevance'. It may be observed that the existing counter-examples use a Model 1 interpretation of the phrase; the proofs appear to use a Model 2 interpretation. Clearly the matter requires further investigation.

ACKNOWLEDGEMENTS

M.E.M. and W.S.C. gratefully acknowledge the support provided by the Committee on Research of the University of California at Berkeley. S.E.R. gratefully acknowledges the opportunity afforded by a visiting appointment at the School of Library and Information Studies of the University of California at Berkeley. This work was supported in part by N.S.F. Grant No. IST-7917566.

REFERENCES

Cooper, W. S. and Huizinga, P. (1981) *The Maximum Entropy Principle and its Application to the Design of Probabilistic Information Retrieval Systems*, Technical Report, School of Library and Information Studies, University of California, Berkeley, California 94720.

Cooper, W. S. and Maron, M. E. (1978) Foundations of probabilistic and utility-theoretic indexing. *Journal of the Association for Computing Machinery 25*, 67-80.

Cox, D. R. (1970) *The Analysis of Binary Data*. London: Methuen.

Croft, W. B. and Harper, D. J. (1979) Using probabilistic models of document retrieval without relevance information. *Journal of Documentation 35*, 285-295.

Gokhale, D. V. and Kullback, S. (1978) *The Information in Contingency Tables*. New York: Marcel Dekker Inc.

Maron, M. E. and Kuhns, J. L. (1960) On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery 3*, 216-244.

Robertson, S. E. (1977a) Progress in Documentation: Theories and models in information retrieval. *Journal of Documentation 33*, 126-148.

Robertson, S. E. (1977b) The probability ranking principle in IR. *Journal of Documentation 33*, 294-304.

Robertson, S. E. and Sparck Jones, K. (1976) Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129-146.

Tribus, M. (1969) *Rational Descriptions, Decisions, and Designs*. Oxford: Pergamon Press.

Van Rijsbergen, C. J. (1979) *Information Retrieval* (Second edition.) London: Butterworths.