THE UNIFIED PROBABILISTIC MODEL FOR IR

S.E. Robertson
Centre for Information Science
The City University
Northampton Square
London EC1V 0HB
U.K.

M.E. Maron
School of Library &
Information Studies
University of California
Berkeley, California 94720
U.S.A.

W.S. Cooper

In this paper, we propose to discuss a probabilistic model for IR which we have recently developed, and which aims to unify the two previous approaches to the problem. The model itself is described in some detail in a recent paper (Robertson, Maron & Cooper, 1982), and the context has been covered by the previous speaker, so the aim of the present paper is to give a brief description of the model and then to pursue some of the implications of the way of looking at IR suggested by the model.

It should perhaps be pointed out that while the framework of the model was laid down during a visit of the present speaker to California, this paper has been written by him after the event, with the considerable disadvantage of some 6000 miles separating him from his co-authors. He should, therefore, be blamed for the worst infelicities in this paper.

Framework for a unified model

The two earlier models (Models 1 and 2 respectively) dealt with situations in which the system possesses data about the individual document in relation to a class of queries, or about the individual query in relation to a class of documents. (If there is data about the individual query in relation to the individual document, then no retrieval system is necessary.) Suppose, therefore, that we have both kinds of information. What kind of model do we need to take account of all the information we have in assessing probability of relevance?

The situation is formally described in terms of the following notation.

A = the class of all uses of the system (queries)

C = the class of all documents in the system

b_k = an individual use

d_m = an individual document

Then the event space with which we are concerned is $A \times C$, and relevance (under the

usual assumptions) is a relation

$R \subseteq A \times C$

that is, $(b_k, d_m) \epsilon R$ if and only if d_m would be judged relevant to b_k.

We also have classes of similar documents/queries, defined by the properties of those documents/queries:

$B \subseteq A$ is a class of similar uses.

$D \subseteq C$ is a class of similar documents.

We may represent the situation diagrammatically as follows: the entire event space is represented by an $A \times C$ matrix; a section of that matrix is the smaller $B \times D$ matrix; a cell of the $B \times D$ matrix is the individual (b_k, d_m) pair.

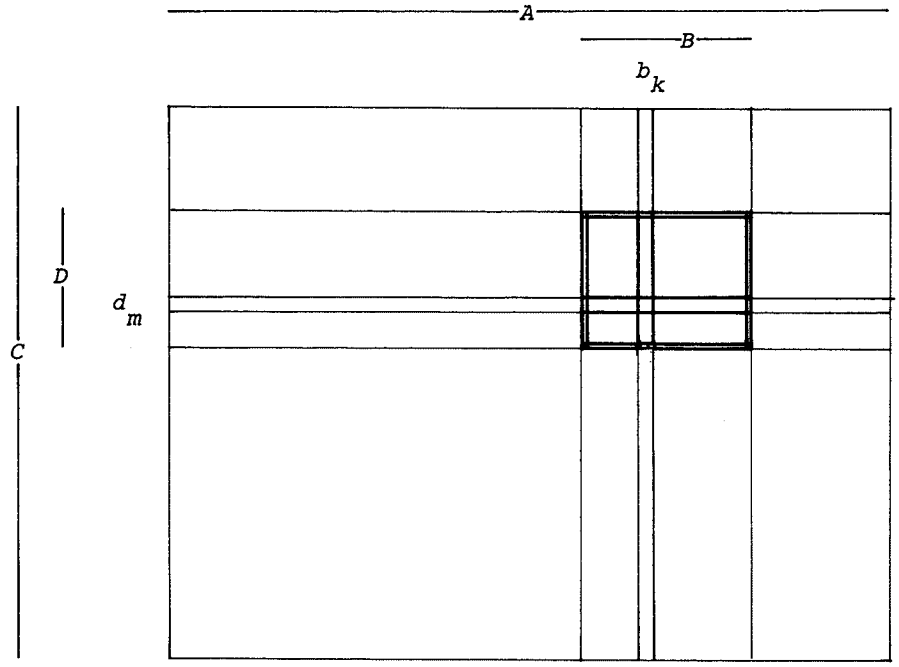

Figure 1: The $B \times D$ matrix in the context of the entire event space.

This formal model carries with it certain underlying concepts, to which we shall return.

Classes and properties

The classes (of documents or queries) used in the model are defined by the properties of these entities. Thus it is assumed that the entities have identifiable properties. In the two earlier models, these properties were refered to as index terms. This

terminology confused the task of unifying the models, as we shall see below.

Whatever these properties are, it is clear that each document or query might possess several of them. Thus the "class of similar documents" (or queries) must be defined as the class of documents (or queries) that possess exactly the same set of properties. This matter is discussed at length in the earlier paper.

Model 3

We are concerned with the B x D matrix. Our supposition, above, was that the system possesses data about the two probabilities:

$P(R|B,d_m)$ (the quantity which figures in Model 1)

and $P(R|b_k,D)$ (the quantity which figures in Model 2)

(This data may take the form of a frequency estimate, or otherwise.) This is essentially marginal information about the B x D matrix; we may also suppose that we have data about

$P(R|B,D)$

We do not, however, have data on whether or not

$(b_k,d_m) \varepsilon R$

What we want is a value for the probability of this event, based on the data that we have. This must of necessity involve a non-interaction model - that is, a model that describes the individual cell probability in terms of marginal values only.

Thus Model 3 consists of a non-interaction model, which specifies $P(R|b_k,d_m)$ (or an estimate of this quantity) in terms of $P(R|B,d_m)$, $P(R|b_k,D)$, and $P(R|B,D)$. The problem is: what should this non-interaction model be? The answer to this question is not as obvious as perhaps one might expect.

Non-interaction models

A non-interaction model implies a set of assumptions of independence between certain events (which may or may not be explicit). We may start by thinking of possible independence assumptions. A pair of assumptions which make fairly obvious candidates are:

$$P(b_k,d_m|R) = P(b_k|R)\ P(d_m|R)$$

and $$P(b_k,d_m|\bar{R}) = P(b_k|\bar{R})\ P(d_m|\bar{R})$$

within the B x D matrix (i.e. assuming all probabilities conditional on B and D). In words this says: given B, D, the events b_k and d_m are independent conditional on R.

These assumptions lead very easily to a simple formula for Model 3. Using odds $O()$ instead of probabilities,

$$\text{i.e.} \qquad O(X) = \frac{P(X)}{1 - P(X)}$$

the formula is

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

This is refered to as the "odds formula".

What is wrong with the odds formula? The answer lies in the nature of the event space and probability measure. Our central assumption is that each pair (b_k,d_m) from the event space A x C has equal probability to start with. From this it follows that (among other things):

$$P(b_k,d_m) = P(b_k)\,P(d_m)$$

Unfortunately this result is not in general compatible with the independence assumptions that we made in deriving the odds formula. The consequence of this is that if we apply the odds formula to all the cells in the B x D matrix, the resulting values are not consistent with the marginal totals with which we started.

Could we abandon our central assumption about the uniformity of the probability measure? This would invalidate (a) any simple estimation of probabilities from frequencies, and (b) any obvious definition of such quantities as $P(b_k|R)$ in terms of the A x C event space (these matters are discussed further in the earlier paper). Thus we do not feel inclined to do so!

Can we think of alternative non-interaction models? In the earlier paper we considered two, a linear logistic model (Cox, 1970) and a maximum entropy model (Cooper & Huizinga, 1982). These turned out to yield the same solution, so the arguments for using this solution are strong. Unfortunately (at least from the point of view of exposition), the solution has to be derived by an iterative method.

The unified model

For Model 3, we assumed that the system possesses all the data that might be used in either Model 1 or Model 2. In general, we need a model that accepts partial information, e.g. information on some individual documents but not others. We therefore proposed in outline a unified model, as follows. We have first to identify a Model 0 which uses the quantity $P(R|B,D)$, when there is no individual information about

according to probability of relevance, where this probability is one of the following:

$P(R|B,D)$

$P(R|B,d_m)$

$P(R|b_k,D)$

$P(R|b_k,d_m)$

depending on what data the system has. The first three quantities may be estimated direct from relevance feedback data, for example, where this data exists, or perhaps from subjective guesswork (as has been proposed for Model 1). When we have all three, then the fourth may be obtained by means of the appropriate non-interaction model.

This is by no means a complete specification of a unified model; we have to consider in addition the situation where we have partial or imprecise (e.g. small-sample) data on any of the probabilities. This problem remains to be tackled.

Conceptual models

So far we have discussed the formal models, including a notation (which implies a certain logical structure) and the actual probabilistic models. But underlying any formal model lies a conceptual structure, perhaps represented only in the language that is used to describe the situation. It is the contention of this paper that there was a particular conceptual structure underlying the earlier models, apparently common to both of them, in which changes have been forced by the development of the unified model. These changes are of significance beyond their application to probabilistic models, and in particular in the common notion of an index language. The remainder of this paper is devoted to an analysis of these changes and their consequences.

Old conceptual model

The conceptual model which, as I understand them, underlay both Models 1 and 2 in their original formulations is easily described diagrammatically:

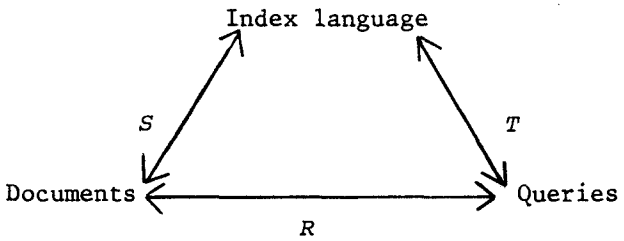

Diagram 1

R is, as before, the relation of relevance;
S is the assignment of terms to documents (indexing);
T is the choice of terms for searching.

There is also a matching function Z, which matches documents and queries via S and T (e.g. by counting the index terms that a particular document and query have in common). Outside the framework of the diagram, the index language is initially defined by means of specified relationships with other entities (e.g. natural language, the state of the subject, etc.)

Model 1 asks the question: "Given T and R, how do we optimize S and Z?" Any external referents of the index language are ignored: in effect, it is assumed to be defined by the choice of query terms T (together with the relevance judgements R).

Model 2 asks the question: "Given S and R, how do we optimize T and Z?" In parallel with Model 1, the index language is assumed to be defined by the choice of indexing terms S (together with the relevance judgements R).

Using the same underlying model, what would be the appropriate question to ask for a unified model? Would it make sense to ask the question: "Given R, how do we optimize S, T and Z?"? It appears that this question is fundamentally unanswerable. The following argument illustrates this point.

Suppose we have two documents d_1 and d_2, two queries q_1 and q_2, and two terms t_a and t_b. Suppose that R is specified as follows: d_1 is relevant to q_1, d_2 is not; d_2 is relevant to q_2, d_1 is not. Suppose further that S and T are not given; we wish to use the information that we have in order to decide on optimal assignments for S and T.

One optimal answer would be to assign t_a only to d_1 and q_1, and t_b only to d_2 and q_2. Another would be to assign t_b only to d_1 and q_1, and t_a only to d_2 and q_2. But there is no way to decide between these alternatives. That is, the searcher cannot in principle know which way round the indexer has chosen to use the terms, and vice versa.

This problem could perhaps be dealt with by taking account of the point mentioned earlier - that the index language is defined by external relationships. But resolving the problem this way would involve building a model which explicitly included these external relationships. Such a model would probably be a great deal more complex, even if feasible, and would involve other entities (such as concepts considered separately from the words used to describe them).

A second way round the problem is to modify the conceptual model described by Diagram 1. This is the course we have pursued.

Modified conceptual model

The first point to be made is that the process of representing documents or queries in terms of an indexing language is often described as a two-stage process. In the case of the document, it has first to be analysed; those aspects or properties of it that are to be indexed must first be recognized, either by a human indexer or by a machine. Thus we may talk of document properties - i.e. those characteristics of the document that may be identified without reference to use or users. Secondly, these properties must be expressed in terms of the index language, which is a device for ensuring that concepts from documents and equivalent concepts from queries are expressed in the same terms. Thus we have the following modified version of Diagram 1.

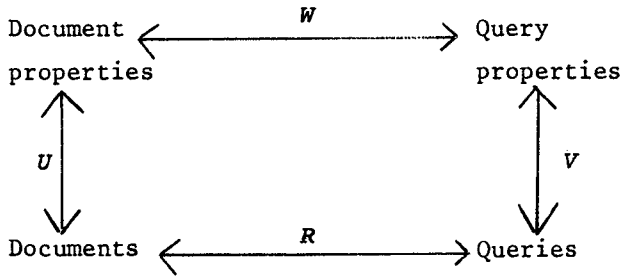

Diagram 2

U is now the recognition/selection of properties of documents by the indexer;
V is the recognition/selection of properties of the need by the user;
W is the matching of document and need properties by the index language.
The matching function *Z* makes use of all the relations *UWV*. We may illustrate this interpretation of information retrieval by means of an example. An index language may say, for instance, "*solar energy* use *solar power*", thus specifying that a document having the property of being about solar energy is to be matched with an enquiry about solar power (or vice versa). A slightly less obvious example would be "*rats* use also *rodents*", indicating that a document about rats is to be matched with a query about rodents - but not necessarily vice versa.

So instead of regarding the index language as a set of entities (*rats*, *rodents*, *solar power*, but not *solar energy*), we see it as a relation between entities, namely document and need properties. These two sets of properties may or may not take the same form.

Now consider some other activities that may well be included in the processes of indexing and search formulation. If the searcher already knows of a particular document in the right area, s/he may use it in searching, either by looking for documents by the same author or by doing a citation search. Either way, the

descriptive of the need; rather, it is a direct prediction by the searcher of some document properties that may be useful. Thus the searcher is specifying document properties not via *V* and *W*, but directly. Similarly, an indexer may specify need properties directly ("Statistics for Social Scientists" is an example.) Thus we end up with Diagram 3.

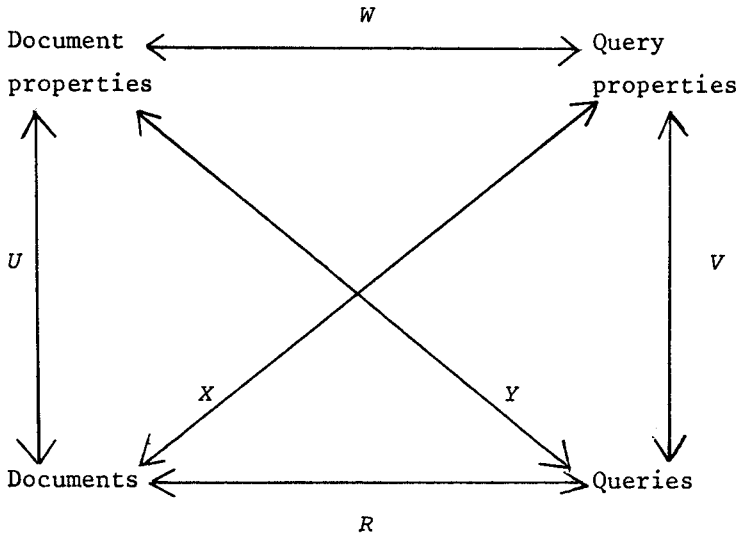

Diagram 3

The matching function can now use any or all of the following combinations: *UWV; UY; XV*.

We now reconsider the probabilistic models in the light of Diagram 3.

Probabilistic models revisited

Model 1 is now seen as seeking to optimize *X* and *Z*, given *R* and *V*, but ignoring document properties and the relations involving them. Similarly, Model 2 seeks to optimize *Y* and *Z*, given *R* and *U*, but ignoring query properties.

Before proceeding, we must distinguish two alternative ways of applying the probabilistic models. Model 1, for example, may be applied by using an indexer's perception of the individual document and the properties of queries, to get the indexer to make probabilistic statements about *X*. Alternatively, the system may (over a period of time) gather data about the relevance judgements *R* and the usage of query terms *V*, and use this data to define *X*. (Both processes may also be used together; the previous speaker has suggested a third.)

In developing the unified model, we assume the latter process of data gathering, although the model should also be applicable to human decision-making at various stages.

In a data-gathering version of Model 1, the system eventually has enough information to define X accurately. However, it must start with some means of retrieving documents about which not enough information is available. The obvious mechanism for this is U, W and the document properties. Thus although in its simplest form Model 1 does not refer to these entities, in practice (in a data-gathering version) it probably uses them to start with.

The unified model, then, looks like this. We assume that documents and queries have properties; the process of identifying or extracting them is outside the scope of the model. Initially, that is all the information we have about them: each document (and each query) is identified only as a member of a class, i.e. those that have identical properties. The index language, as embodied in W, provides the prior evidence about the probability of relevance of each class of documents to each class of needs.

Then we gather data, about relevance judgements made on individual documents in respect of classes of needs, and those made on individual needs in respect of classes of documents; that is, we gather direct data on X and Y. Subsequent retrieval acts depend on as much data as is available, i.e. use the X and/or Y data where possible, but reverting to W where it is not possible.

The four specific models that are incorporated into the unified model are now identified very simply: Model 0 is used if there is no data for X or Y; Model 1 if there is data for X but not Y; Model 2 if there is data for Y but not X; and the new Model 3 if there is data for both.

Conclusions

The two previous probabilistic models of information retrieval, which seemed to be in some sense incompatible, can now be regarded as two complementary parts of a unified model. The new Model 3, which is derived in the framework of the unified model from a combination of Models 1 and 2, makes use of relevance feedback information from the individual user about other documents, and from other users about the individual document.

A necessary consequence of the unification of the two models was a reconsideration of the underlying conceptual basis, and in particular of the role of the index language.

Acknowledgements

This work was supported in part by NSF Grant No. IST-8113213.

References

Cooper, W.S. and Huizinga, P. (1982). The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Development 1*, 99-112.

Cox, D.R. (1970) *The analysis of binary data*. London: Methuen.

Robertson, S.E., Maron, M.E. and Cooper, W.S. (1982). Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development 1*, 1-21.