1. Introduction

1.1 Background

Since the early sixties, attempts have been made to use statistical methods in information retrieval - that is, to devise techniques of retrieval which make use of statistical information in order to gain benefits in effectiveness or efficiency. Many such techniques have been proposed, and some tested, in which statistical information is used in various ways and for various components of the system. Examples are various forms of automatic indexing, term or document clustering, and automatic profile construction or modification. Recent work has tended to concentrate on so-called probabilistic models of IR. In these models, an attempt is made to discover techniques that are in some sense optimal, that is techniques which use the statistical information in the best possible way, in terms of accepted performance criteria. In particular, there are a number of models which give explicit rules for certain operations in IR, on the basis of optimizing retrieval effectiveness, as traditionally measured by recall and precision or related parameters.

There is now a considerable body of work, here and in the States, on the use of such models. This work is both theoretical and experimental, and concerns the use of models (a) for term weights based on relevance feedback, (b) to provide test yardsticks, (c) for term weights based on prior information, and (d) for automatic indexing.

The most consistently successful model, here called the Robertson/ Sparck Jones model, is also the simplest. There are, however, some problems that have been indentified with this model, notably in the area of estimation. In the more complex models, such as those involving dependence or within-document frequencies, these problems take on major proportions, and evidently overcome any substantial benefit that might be derived from the models.

The immediate aim of the project was to bring statistical expertise to bear on these problems, in order to discover whether any existing statistical techniques could be applied to them, or whether new techniques could be devised. Ultimately, this project is seen as part of a continuing effort both to improve the performance attainable in IR systems, and to increase our understanding of information retrieval processes.

1.2 The state of art

Probabilistic models for IR are generally concerned with the probability that a document will be relevant to a query, given that it contains (is indexed by) a particular combination of index terms. It can be shown (under some assumptions) that ranking the documents in order of their probability of relevance given some information about them leads to the best retrieval performance that is possible with that information (Robertson, 1977).

1.2.1 The Robertson/Sparck Jones Model

The simplest probabilistic model for searching is that proposed by Robertson and Sparck Jones (1976). The assumptions on which the model is based are (a) that the indexing is binary or dichotomous (terms are either present or absent), and (b) that within the set of relevant documents, and also within the set of non-relevant documents, occurrences of different terms are statistically independent of each other. These latter assumptions are not regarded as describing the true situation - rather they are simplifying assumptions made in order to render the mathematical problem tractable. Because of these assumptions, the Robertson/Sparck Jones model has also been called the Binary Independence model.

Given these assumptions, it is possible to prove that ranking the documents in order of their probability of relevance can be achieved by assigning weights to the terms according to a certain formula, and performing ordinary weighted retrieval. In other words, retrieval performance will be optimized if this particular weighting scheme is used. The formula is :

$$w = \log \frac{p(1 - q)}{q(1 - p)}$$

where p is the probability of the term occurring in a relevant document, q is the probability of the term occurring in a non-relevant document, and w is the weight for the term in question.

The probabilities p and q clearly have to be estimated in some way. In the case of the use of the model for relevance feedback, the estimate for w becomes

$$\hat{w} = \log \frac{(r + 0.5)\ (N - R - n + r + 0.5)}{(R - r + 0.5)\ (n - r + 0.5)} \qquad (1)$$

where R is the total number of known relevant documents
r is the number of known relevant documents containing the term
n is the total number of documents containing the term
N is the size of the collection.

Some comments on this estimation formula, and in particular on the use of N and n rather than known non-relevant documents and on the "0.5" component, are made in section 2.41.

This model has been extensively tested by Sparck Jones (1979a, 1979b, 1980 ; Sparck Jones and Webster, 1980). It has been shown to perform well under a wide range of conditions, including that of very little relevance information.

However, some problems have emerged. It appears that the formula used for estimating the relevance weight may not be entirely appropriate for the small sample sizes involved, and that this problem affects the performance attainable. Harper and van Rijsbergen (1978) suggest an alternative formula which, although having no obvious theoretical justification, under some conditions outperforms the Robertson/Sparck Jones one.

Further theoretical problems can be discerned in the estimation area. Firstly, the samples used to estimate the various parameters are of necessity biased. This problem has been completely ignored in previous work (apart from a brief discussion in the original paper), and deserves to be reconsidered. Secondly, this bias may be affected by the form of the initial search. All experiments so far have used crude coordination levels for the initial search; one would obviously think of using the best possible non-feedback strategy initially, but this might in fact exacerbate the problem of bias.

1.2.2 Collection frequency

A topic of current concern, related to the Binary Independence Model, is the use of raw collection frequency information to predict term value. It has long been known that frequency is a useful predictor of value; however, there are two competing models of the relationship. One (due to Sparck Jones) assumes a monotonic relation, so that the best terms are the least frequent ones; the other (due to Salton and others) postulates an optimal frequency. An interesting question arises : can a term-frequency model be developed from the Binary Independence Model, and if so, which form of relationship does it suggest ?

Croft and Harper (1979) develop a very simple model on these lines, which supports the Sparck Jones monotonic function. An earlier model with the same results is proposed by Robertson (1976). Salton and others, however, have several models generating their peaked relationship; in particular, Yu, Lam and Salton (1982) have a model based on the Binary Independence Model.

It seems appropriate at this point to look for a unification of the collection frequency weights with the probabilistic models.

In particular, it seems appropriate to include a specifically collection-frequency dependent prior distribution in an estimation formula (in effect, the "*0.5*" formula contains such a distribution implicitly), so that experiments can be done on the appropriate form of prior distribution.

Many experiments described in section 3 make use (in one way or another) of the Sparck Jones collection frequency weights (Sparck Jones, 1972; Sparck Jones and Webster, 1977), namely :

$$w = \log N/n$$

with the same notation as in section 1.2.1.

1.2.3 Term dependence and other statistical properties

There is a considerable amount of work on the use of term-dependence information in probabilistic models. The original idea was that the dependence between terms in the relevance set and non-relevance set of documents should be modelled explicitly (and the appropriate parameters should be estimated from relevance feedback data). This idea has not, so far, generated performance improvements (Harper 1980), probably because of the difficulty of obtaining adequate estimates of the large numbers of parameters involved from the small amount of data typically available. A second problem is the lack of a systematic series of models, starting with independence models and including dependencies one at a time (or at least in a graded manner).

A second mode of use of dependence information is to look at dependencies in the whole collection of documents (not differentiating between relevant and non-relevent), in order to suggest new terms for inclusion in the query (i.e. those terms that are closely associated with query terms). This procedure is known as Query Expansion; the resulting expanded query may be weighted according to the Binary Independence Model (van Rijsbergen, Harper and Porter, 1981). This procedure has produced some performance benefits, though not consistently over different collections.

Another statistical property which may be of use in a probabilistic model is the data on within-document frequencies of terms. Probabilistic models have been proposed which include within-document frequencies, but so far without generating significant performance improvements (Robertson, van Rijsbergen and Porter, 1981; Croft, 1981).

1.3 Related statistical work

The present project starts from the premise that the models and methods currently in use or under investigation in probabilistic IR could benefit from a re-appraisal from a statistical point of view. More specifically, we hoped that it might be possible to make use of models or methods developed for other applications, or at least to learn something from work in other areas.

We started, in fact, with one such area in mind. The area is medical diagnostics ; one can draw a clear formal analogy between the dianostic problem and the IR problem, as follows :

	Medical Diagnostics	Information Retrieval
Aim	To determine the underlying disease or condition, or appropriate action.	To predict relevance to a user's need (as judged by user)
Method	To diagnose the condition on the basis of available evidence.	To rank documents in order of probable relevance, on the basis of available evidence.
Evidence	Symptoms and observations; relationship of symptoms to condition must be established by means of a training set.	Terms; relation of query or other terms to relevance must be established by means of a training set.

Training set	Sample of people, perhaps selected on the basis of possession of certain symptoms.	Sample of documents selected on the basis of possession of certain query terms.

The analogy is good but not perfect; for example, the IR system does not seek to make a decision on which are the relevant documents, but to rank the documents in order to facilitate the user's decision process. Nevertheless, it seems strong enough to warrant an examination of the methods used in statistical medical diagnosis.

It turns out that work in the medical diagnosis area has gone through stages similar to those apparent in IR. That is, independence models were tried and found useful, but attempts to improve on those results by including dependencies were not very successful (Titterington et al., 1981). There was, however, one class of models which seemed to represent an advance on anything in use in IR : namely, the class of *linear logistic* models. Specific reasons for the use of this class of models in diagnosis, which all seemed relevant to IR, are :

(a) that the class includes "independence" models as well as models which allow for interactions between symptoms;

(b) that the "independence" models are very much less restricted than the simple probabilistic models such as Robertson/Sparck Jones;

(c) that the models and their associated estimation procedures do not assume that the estimation sample (the training set) is a random sample of the population (Dawid, 1976)

Further, we failed to find any other areas where work on similar models was going on. Therefore most of this project has been devoted to developing and using the linear logistic models for information retrieval.

1.4 The present project

In the light of previous section, the project can be divided into three areas :

(a) Development of the model and methods for applying it to IR. This was largely taken from the medical work, but required a significant amount of progamming work on the estimation method to be applicable to IR.

(b) Running experiments with the new model on existing test collections.

(c) Further theoretical development, with a view to answering questions about when new terms or interactions should be included in the model.

Chapters 2,3 and 4 correspond approximately to these three areas.