Information Processing & Management Vol. 26, No. 3, pp. 371–382, 1990
Printed in Great Britain.

0306-4573/90 $3.00 + .00

A COMBINATION OF EXPERT OPINION APPROACH TO PROBABILISTIC INFORMATION RETRIEVAL, PART 1: THE CONCEPTUAL MODEL

PAUL THOMPSON
College of Information Studies, Drexel University, Philadelphia, Pennsylvania 19104, U.S.A.

(*Received* 28 *April* 1988; *accepted in final form* 7 *June* 1989)

Abstract—This paper presents a new version of Robertson, Maron, and Cooper's probabilistic information retrieval model, Model 3. The new version of Model 3, based on the statistical technique of the combination of expert opinion, attempts to overcome difficulties with the previous version of Model 3: Those faced by most probabilistic retrieval models including various independence and small sample problems and those specific to Model 3. Model 3 seeks to combine two earlier models of retrieval, but how to do so is unresolved. The present model shows how that combination can occur. Furthermore, by advocating the use of indexer and user subjective probabilities and determination of dependent terms, it shows a way of dealing with the problems of small samples and independence/dependence of index terms.

I. INTRODUCTION

This paper introduces a new model of probabilistic information retrieval (PIR) based on the statistical technique of the combination of expert opinion. This model can be viewed as a generalization of Robertson, Maron, and Cooper's (RMC) unified PIR model, Model 3 [1]. Part 2 will provide the mathematical details of what is presented here conceptually. PIR models go back as far as Maron and Kuhns [2]. In the last ten years or so information retrieval (IR) theoreticians have become increasingly interested in probabilistic models, e.g., Bookstein [3,4], Cooper and Huizinga [5], Cooper and Maron [6], Croft and Harper [7], Harter [8,9], Radecki [10], Robertson and Sparck Jones [11], Yu and Salton [12]. However, recent experimental work in PIR has led to less satisfactory results than expected. For instance, on the basis of their experiments Smeaton and van Rijsbergen [13] observed that due to the small sample of relevant documents from an initial retrieval, the problem of estimating term probabilities was much more serious than previously recognized.

It may be that the small sample problem, and others to be discussed below, cannot be resolved successfully within the current paradigm of IR, and in particular PIR, research. Smith [14] has characterized this paradigm as that of machine intelligence, i.e., emphasis has been on fully automatic retrieval. For example, goals of such research have included automatic indexing and classifying of documents—tasks ordinarily done using human intelligence. She contrasts this with an alternative paradigm: machine-aided intelligence. The development of interactive computing makes an approach to PIR drawing on the strengths of humans as well as those of the computer desirable. An indexer examining a document, or a user reflecting upon her information need (for convenience in what follows users will arbitrarily be feminine and indexers masculine), is capable of much deeper understanding of document and need than can be simulated by a computer. With interactive computer assistance in elicitation and updating based on users' relevance judgments this depth of understanding should provide more accurate probability estimates than the computer alone could. The computer could assist by providing various statistics and other information about documents or past searches as well as by debiasing and calibrating estimates. The PIR model of this paper using indexers' and users' subjective probabilities provides a machine-aided intelligence approach to PIR which may offer a way of largely overcoming obstacles faced by machine-intelligence PIR models. The issues concerning the use of subjective relevance judgments have been discussed by a number of writers [15–18].

2. PROBABILISTIC MODELS OF INFORMATION RETRIEVAL

Document retrieval has often been viewed as a simple, deterministic matching operation. An indexer assigns index terms from a thesaurus to a document. A user of a retrieval system formulates a query by selecting search terms from the same thesaurus and combining them with Boolean operators such as AND, OR, or NOT. A document is retrieved whenever its indexing satisfies the logic of the query, otherwise it is not retrieved. Such a procedure ignores the inherent uncertainty of document retrieval. A document that a user would in fact judge relevant might not have been indexed by the same terms that the user would select for her query. Two users making identical queries which retrieve the same documents may differ greatly in their judgments of which documents are relevant. The fact that a given term has been assigned to a document or used in a query cannot determine that a particular user will judge a certain document relevant. Rather, these term assignments should be viewed as probabilistic clues which can be used in an attempt to predict relevance. Accordingly, a probabilistic approach to document retrieval is needed.

The Probability Ranking Principle, originally stated by W.S. Cooper, provides the rationale for PIR [19]:

> If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

It seems reasonable to add that as much data, or evidence, as possible should be made available to the retrieval system, provided, of course, that the data is sufficiently valid and reliable. Some PIR models use only one kind of probabilistic evidence. The model presented in this paper combines several kinds of evidence, as does RMC's Model 3 [1] on which it is based.

3. ROBERTSON, MARON, AND COOPER'S MODEL 3

According to Maron and Kuhns' [2] probabilistic theory of indexing, the task of indexing can be described as follows: Given a specific document, call it D, then for every term the job of the indexer is to estimate subjectively the probability that if a searcher were to want D (i.e., judge the document relevant) and if she were using only a single term as her search query, then she would select that term for her search query. Thus an indexer estimates a probability of relevance for a particular document for the class of users using that search term. This approach has come to be called Model 1.

Later, Robertson and Sparck Jones [11] developed an alternative approach to PIR. They considered that index terms were assigned without weights, but that probability entered into the search formulation stage. Retrieval took place in several iterations. After each iteration the user judged which retrieved documents were relevant, i.e., documents that the user would want. This feedback allowed calculation of a probability of relevance for each term according to its distribution in relevant and nonrelevant documents. Determined on the basis of a particular user's relevance judgments, these were probabilities of relevance for a particular user and the class of documents indexed by a given term. Thus, this approach, which came to be known as Model 2, is a probabilistic theory of searching.

RMC's [1] unification of Models 1 and 2, Model 3, was achieved with the realization that the relevance relation involved two sets of entities, i.e., users and documents, the elements of each of which could be considered to have properties. Thus instead of just having an indexer predict which search terms a user would select were she to judge a given document relevant, as in Model 1, or using a user's relevance judgments to predict whether a document having a certain term would be relevant, as in Model 2, it was possible to do both. An indexer would assign (or not assign) terms from a thesaurus of document prop-

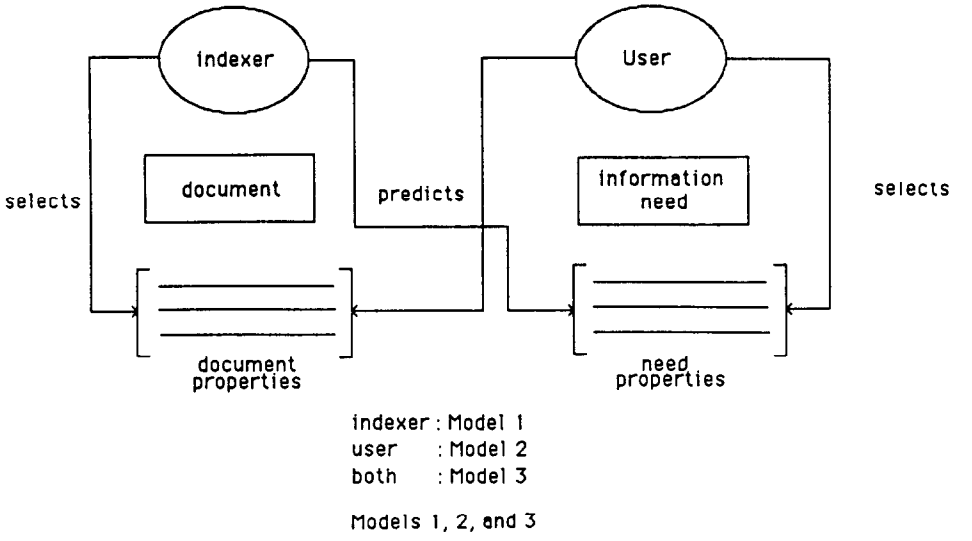

Fig. 1.

erty terms. At the same time he would predict probabilities of relevance for the document for users with various properties. The user would describe her need by selecting terms from a thesaurus of user property terms and predict, at least indirectly through relevance judgments, probabilities of relevance for documents having various properties already selected by indexers. The Models 1 and 2 probabilities would then be combined to give a final Model 3 probability. Figure 1 shows a schematic view of Models 1 and 2 with separate thesauri of property terms. Their combination results in Model 3.

Although as presented here document properties can be thought of as index terms which are assigned (or not assigned) to a document and user properties as search terms which are selected (or not selected) by a user, document and user properties can be viewed in a more general fashion. Document properties could include any features that a document might have, e.g., frequencies of words in the text, or co-citation patterns with other documents. Similarly, user properties might include other features of the user or the information need, e.g., the user's organizational affiliation.

4. PROBLEMS WITH MODEL 3

Model 3 faces two sorts of difficulties. The first include those of PIR models in general, i.e., independence/dependence, small samples, and the validity of the probabilities used. Validity is most commonly defined, in the context of social science research, as the extent to which one is actually measuring what one believes oneself to be measuring [20, p. 457]. As an example of an independence problem, term probabilities in a request are commonly assumed to be independent given relevance or non-relevance, though this assumption is widely recognized as being unjustified. Various solutions have been proposed to allow for some dependence through use of the maximum entropy principle, maximum spanning trees, or more complete expansions such as the Bahadur-Lazarfeld Expansion. Unfortunately, recent experimental research has tended to show that the more theoretically sound the principle used to accommodate dependence, the worse the experimental results [13,21]. This has been attributed to the small sample problem when such techniques are based on the small number of documents found relevant by a user in an initial retrieved set. Various approaches have been suggested to overcome this problem [7,22,23]. Another cause for poor performance may be the way probabilities of relevance are derived, aside from problems of dependence. Typically they are based on statistics taken from patterns

of postings of documents to terms, counts of words in documents and so on. These approaches, though sometimes plausible sounding, have not been completely verified.

Model 3 also has more specific problems. These mainly concern how Models 1 and 2 are to be combined to obtain Model 3. Most ways of combination suggested by RMC rely on an additional model, Model 0, involving the probability that a randomly selected document from the class of documents with identical indexing (with respect to terms specified by the user) will be judged relevant by a user randomly selected from the class of users who have described their information need in an identical way. The PIR system is to obtain data with which to estimate these probabilities from relevance feedback. Unfortunately in many cases there may be little or no data on which to base these estimates, i.e., there may be a severe small sample problem.

Aside from problems with Model 0, RMC make several implicit assumptions which seem questionable: That Models 1 and 2 are independent of each other; that each model should be given equal weight; and that the probabilities are calibrated, i.e., accurate. Finally, RMC, do not explicitly show how relevance feedback is to be incorporated.

As will be shown below, the statistical technique of combination of expert opinion, or CEO, provides a solution to the problem of how to combine Models 1 and 2. Furthermore, by using indexers' and users' subjective probabilities and automatic Bayesian updating, i.e., revision of probabilities based on additional information, in this case relevance judgments, the model of this paper, which will be referred to as CEO Model 3, shows how the independence/dependence, small samples, and validity problems might be largely overcome as well—if these subjective probabilities are sufficiently reliable. Thus, CEO Model 3 provides a good example of what Spiegelhalter [24] has called a "coherence machine," i.e., a Bayesian system which begins with subjective judgments and learns from subsequent data.

5. USE OF SUBJECTIVE PROBABILITIES

Using indexers' or users' subjective probabilities, or utilities, has been considered in the IR literature, e.g., [2,17,25], but the psychological issues involved seem not to have been given serious attention until recently [26]. The fundamental question with respect to their use is: Can humans make reasonably accurate probability estimates? Subquestions related to this are: How well do people estimate probabilities without any special training?; Is training effective?; Are there systematic biases in human estimation?; If there are, can these biases be corrected and, if so, how?

An extensive review of the psychological literature (see Kahneman *et al.* [27] for an introduction to this literature) suggests that although human probability estimation is a difficult task and subject to biases, at least some people, such as meteorologists, can do it well. Training and feedback can improve estimation. Debiasing techniques are also being developed. Furthermore, there is some evidence that the poor results from some studies of probability estimation may be largely due to the artificial nature of experimental tasks and that in real situations people do better. Probability estimation performance also seems to be quite task dependent. Thus, to determine how effectively indexers or users could make estimates, experiments must be conducted in the context of information retrieval.

If, then, it were feasible to use indexers' and users' subjective probabilities, the difficulties facing PIR mentioned above, i.e., independence/dependence, small samples, and validity, could be met as follows. An indexer ordinarily has subject area expertise and a user, of course, is intimately concerned with her own information need. Their subjective probabilities thus have validity. While there is no small sample problem as such with subjective probabilities, i.e., a person can have a degree of belief in the occurrence of some event without sampling, the related problem of reliability is a concern, but considerations just referred to from the psychological literature indicate that this may not be an insurmountable problem. Furthermore initial indexer estimates could be improved through calibration, i.e., adjustment of estimates based on past performance, and updating based on relevance judgments.

Use of subjective probabilities may also help resolve the independence/dependence

problem. In behavioral decision theory application has been made of research on configural cue utilization which indicates that people can detect dependencies between cues [28,29]. Thus if it can be assumed that there are certain gross dependencies between some terms with respect to a particular document or use, but that among other terms approximate independence holds, then these findings imply that indexers or users may be able to group dependent terms into complexes and make estimates with respect to these complexes as easily as for a single term and that final probabilities calculated on this basis will be considerably more accurate than would be the case assuming independence. This idea has been anticipated in the medical diagnosis field by Gustafson and his colleagues who conducted experiments testing the effectiveness of different methods of determining complexes of dependent terms [30,31]. More recently, in a PIR setting, Croft [32] has experimented with determining dependent terms through Boolean structure (terms connected by AND in disjunctive normal form are assumed to be dependent) or through user specification of important phrases (i.e., dependent terms) in natural language querying. In his approach in a group of n dependent terms possible dependencies among subgroups were ignored. Over three test collections at most, an average of four groups of dependent terms were found; and, at least with the natural language queries, using the dependence information improved precision at standard recall levels. Bookstein [33] has also developed a similar model using Tague and Nelson's notion of hyperterms [34]. Losee and Bookstein [35] report experimental results showing that the use of hyperterms can improve retrieval.

Now, how are these subjective probabilities to be combined to produce Model 3, and how is the final Model 3 probability to be interpreted? To answer these questions it is first necessary to discuss CEO.

6. COMBINATION OF EXPERT OPINION

Various ways of combining probabilities of different experts, or models, have been discussed in the statistical literature. A good review is provided by Genest and Zidek [36]. There are two general situations. In the group consensus problem several individuals, each with a probability (or distribution) for some event or parameter θ, wish to combine their probabilities to obtain a group probability. The simplest way would be to take a weighted average of their individual probabilities. The weight given each individual's probability would reflect how expert that individual was relative to the others. This approach suffers from several theoretical weaknesses including how to determine the weights in other than an ad hoc way and how to account for dependence among the individuals' probabilities [36].

The second general methodology of combining opinion is often called the combination of expert opinion (CEO). The usual Bayesian formulation of the CEO problem, e.g., Winkler [37–39], Morris [40–43], Lindley *et al.* [44], Lindley [45,46], runs as follows: A decision maker is interested in some parameter or event θ; he or she has a prior, or initial, distribution or probability for θ which he or she revises upon consulting n experts, each with his or her own distribution or probability for θ. To effect this revision the decision maker must somehow assess the relative expertise of the experts and their interdependence both with each other and the decision maker. The experts' distributions are considered as data by the decision maker which he or she uses to update his or her prior distribution according to Bayes' theorem. For example in the case of one expert if the decision maker's prior distribution is $p(\theta|H)$, where H represents whatever information the decision maker has before consulting the expert, and $t^{(n)}$ represents a finite set of numbers characterizing the expert's distribution, e.g., sufficient statistics (i.e., numbers which completely describe a distribution such as the mean and variance of a normal distribution), then

$$p(\theta|t^{(n)},H) \propto p(t^{(n)}|\theta,H)p(\theta|H) \qquad \text{or for } k \text{ experts,}$$

$$p(\theta|T_1,\ldots,T_k,H) \propto p(T_1,\ldots,T_k|\theta,H)p(\theta|H) \tag{1}$$

where T_i represents the vector summarizing the ith expert's distribution and $\propto$ indicates 'is proportional to' [45]. In CEO Model 3, for example, T_1 would represent the mean and

variance provided by Model 1. In this paper $p(.)$ will stand for a probability density function; $P(.)$ for a probability function.

As a simple example of CEO involving a single expert and probabilities rather than densities, consider a person interested in whether or not it will rain tomorrow. Let A be the event of rain tomorrow. $P(A|H)$ will be the person's initial probability that it will rain tomorrow, where as before H represents background information. For the remainder of this example H will be omitted, but all the person's probabilities are understood to be conditional on H. If the person receives some new evidence E bearing on the probability of its raining tomorrow, then from Bayes' theorem

$$P(A|E) = \frac{P(A)P(E|A)}{P(E)}.$$

These are all subjective probabilities of the person. $P(E|A)$, called the likelihood of A, is the person's probability for the evidence given that it will rain tomorrow. $P(E)$ can also be written as

$$P(E) = P(E|A)P(A) + P(E|\bar{A})P(\bar{A})$$

where $\bar{A}$ is 'not A' and so

$$P(A|E) = \frac{P(A)P(E|A)}{P(E|A)P(A) + P(E|\bar{A})P(\bar{A})}.$$

Thus, if the person assesses his or her probabilities for the evidence both given that it will and will not rain, then it is possible to calculate what that person's revised, or posterior, probability should be for rain tomorrow given the evidence, i.e., $P(A|E)$, since now $P(A)$, $P(E|A)$, and $P(E|\bar{A})$ have all been assessed and $P(\bar{A}) = 1 - P(A)$. This is merely a simple application of Bayes' theorem.

So far the nature of E has not been specified. Since we are explaining CEO, let E be an expert's opinion. To be precise let E be the observation that a certain meteorologist says that the probability of rain tomorrow is .7. Then $P(E|A)$ is the person's probability that the meteorologist would say .7 when in fact it will rain tomorrow, while $P(E|\bar{A})$ is the person's probability that the meteorologist would say .7 when it will not. The meteorologist's opinion is thus treated like any other evidence and enables the person to update his or her prior probability for rain tomorrow via Bayes' theorem. $P(E|A)$ and $P(E|\bar{A})$ allow his or her expression of the meteorologist's expertise. If the person thinks that the meteorologist is very competent, then the person will consider it unlikely that the meteorologist will say that the probability of rain is .7 when in fact it will not rain. Thus $P(E|\bar{A})$ will be low. Of course the person might also think it unlikely that the meteorologist would say that the probability was .7 if it will rain. Since the person has a high regard for the meteorologist's abilities, he or she might more likely expect the meteorologist to give the probability of rain as .8 or .9. Still $P(E|A)$ will be greater than $P(E|\bar{A})$.

7. CEO APPLIED TO THE UNIFIED MODEL OF PIR

CEO Model 3 follows the general CEO model of Lindley [45]. The PIR system itself is the decision maker. In the CEO literature the decision maker is always human. Nonetheless if a parameter θ can be found (see below) along with a reasonable way to give the system a prior distribution for θ and to make sense of the system's opinion of the experts, then everything else follows automatically from Bayes' theorem. Thus no harm is done by making the PIR system the decision maker. Roughly speaking the experts are Model 1, or the indexers, Model 2, or the user, and a feedback expert, Model 2*. The actual situation, to be explained below, is somewhat more complicated. Unlike the case with RMC, each expert will provide a probability distribution rather than a probability. Thus rather than the PIR

system, as decision maker, determining its probability that, say, Model 1 would give a certain probability as evidence, it would determine its distribution for Model 1's giving a certain mean and standard deviation. There are two reasons for preferring distributions. First it allows more flexible expression of uncertainty. If one wants to elicit an indexer's opinion on the proportion of users with needs of a certain type (e.g., having use property j) who will judge the document under consideration relevant, then it is misleadingly precise for him to give, e.g., an estimate of .75. The second reason to prefer distributions is that they can be more readily updated when relevance judgments become available.

The Models 1 and 2 expert distributions will themselves result from lower level CEOs. The CEO already mentioned, involving the combination of Models 1 and 2 to obtain Model 3, will be referred to as the upper level CEO. At the lower level each indexer will be considered as a multiple expert—a separate expert in the use of each user-need property; while the user will be seen as a multiple expert with respect to each document property. Thus the upper-level Models 1 and 2 experts will be lower-level decision makers who consult Models 1 and 2 term experts.

As a simple example of how these levels of CEO work, consider a particular document, D, which an indexer has identified as having the properties "information retrieval theory" and "Boolean operators." The indexer has also given his subjective probability distributions for the chances that a user with need properties "probabilistic information retrieval" or "information retrieval evaluation" will judge D relevant. Now assume that a user having both of these need properties submits a query and that she wants documents with the document properties mentioned above and thus gives her subjective probability distributions for the chances that she will want a document with properties "information retrieval theory" or "Boolean operators." The user's distributions involving the individual document properties are lower level Model 2 expert opinions which the PIR system, as Model 2 decision maker, combines to obtain the Model 2 distribution. In the same way the indexer's distributions for individual need properties are lower level Model 1 expert opinions which the PIR system, as Model 1 decision maker, combines to obtain the Model 1 distribution. Finally, upper level CEO takes place when the PIR system, as upper level decision maker, combines the Models 1 and 2 distributions to obtain its final Model 3 distribution.

A surprisingly tricky question must now be raised. What is θ, the parameter of interest to the decision maker? One wants to say that it is the probability of relevance of a particular document for a particular user-need as do RMC, but this will not do. If the experts were providing probabilities this would be acceptable, but they are giving distributions and it makes no sense, at least according to some schools of thought, to speak of probability distributions of probabilities. Instead experts will give their distributions for the long-run relative frequency, or proportion, of times that a user of a certain description will judge a document of a certain description relevant. This sequence of judgments forms a Bernoulli sequence, i.e., a succession of successes or failures, like the flipping of a coin for heads or tails. In Bayesian terminology the long-run relative frequency of successes is referred to as a chance.

At first glance it may seem difficult to imagine how the θs estimated by different experts, or the system itself as decision maker, could all be the same θ since it would seem that different experts are considering different sequences. For example, an indexer making a Model 1 estimate of the proportion of users using a given index term, who will judge the document that he is now indexing to be relevant, is considering a sequence of pairings of that document with all users using that index term, while a user making a Model 2 judgment of the proportion of documents identified by a certain document property is considering a sequence of pairings of herself with documents having that property. Nevertheless, it can be shown that there is a common subsequence underlying each expert's sequence as well as that of each decision maker. Since as far as each expert or decision maker is concerned the proportion of successes in any subsequence of the sequence being considering would be the same as the proportion in the sequence as a whole, each expert or decision maker can be said to be estimating θ for the common subsequence. A fuller discussion and mathematical demonstration of this claim is given by Thompson [26].

8. EXPERTS, DECISION MAKERS, AND THEIR DISTRIBUTIONS

Having gotten some insight into the estimation of θ, let us examine the experts, decision makers (remember that Models 1 and 2 are also lower-level decision makers), and their distributions for θ. The Models 1 and 2 experts share many similarities. Each is basically human, the indexer or user, respectively. How, then, can they be considered multiple experts in the use of each term, as mentioned above? Lindley *et al.* [44] discuss the following as a CEO task. A person is interested in introspecting his or her distribution for an event or parameter θ. Instead of making a direct assessment, the person might adopt an initial probability based on some preliminary information and then consider the problem from various perspectives, giving a conditional likelihood for θ based on each perspective. Then using Bayes' theory a revised probability could be obtained by using these likelihoods to update the initial probability. Similarly, one can think of the indexer as assessing probability of relevance for a document based on the perspectives of the user's need being expressed in terms of various properties. The indexer would also have an initial probability, but the analogy breaks down because the indexer never actually updates his probability. An updated probability can only be obtained when a search is underway and a user has specified her need properties. In forming its prior distribution the Model 1 decision maker estimates a probability distribution for the chance in a Bernoulli sequence of all use-document pairs involving this particular document. Since the unaided indexer might find this probability difficult to assess, the system could provide data to assist him, e.g., the frequencies with which various user-need properties were specified in searches, or the indexer might merely rate the document on some scale, e.g., that its prior probability of relevance was either low, medium, or high and then the system with its data would convert this to a distribution. An analogous analysis can be made for the user. These considerations lead to the conclusion that the Models 1 and 2 experts really involve the PIR system as well as the indexer, or user.

Indexers' lower level distributions will be updated by relevance judgments. A Model 1 term expert when indexing a document is estimating the proportion of users with a certain need property, i.e., using a certain search term, that will judge the document relevant. As judgments arrive this estimate can be refined. Since the proportion estimated by the indexer ranges from 0 to 1, the beta distribution is a natural one to use. It can easily be updated with each relevance judgment [47], whereas a normal distribution, for example, could not be. Graphically, the beta distribution can take many shapes, and is thus capable of expressing a wide range of opinion. Schaefer and Borcherding [48] report on a training experiment which indicates that, although initially people have difficulty working with beta distributions, eventually they can become proficient in their use. It seems, then, that indexers could use beta distributions. Moreover improved elicitation techniques for beta distributions are being developed [49].

The Model 2 expert differs from the Model 1 expert in several respects. Users would not need to use beta distributions, since their distributions will not be automatically updated. In fact it may not even be necessary to directly elicit any distribution from the user. It is realistic to assume that errors in probability assessment are normally distributed with constant variance once transformed to log-odds [44]. This does not mean that all distributions will have the same variance, but that assessments of variance, once expressed in terms of log-odds, are just as accurate regardless of the assessed parameter's location. Transformed back to the probability scale this means that midrange probabilities are more difficult to estimate than those nearer 0 or 1. Thus, especially if debiasing techniques had been applied, it would make sense to think of the error distribution of the log-odds of estimates of probability of success in a PIR Bernoulli sequence as being normally distributed. Instead of eliciting a distribution from a user, a probability could be obtained (which would be interpreted as a mean) along with some indication from the user of her confidence in that probability. Perhaps she could be given a choice of three or four levels of confidence which could be interpreted as a standard deviation once conversion to log-odds took place. The second main difference is that the user, unlike the indexer, will be interacting with the PIR system during retrieval, thus allowing subjective user updating. After

the user has seen document surrogates from a given iteration, she will have more information and so can revise her original assessments. If the user has been consistently over- or underestimating term probabilities, the system could inform the user so that she could consider this when revising her initial estimates. As with Model 1, dependence will be largely eliminated through grouping of dependent terms into complexes.

Model 2*, the feedback expert, will follow the general Model 2 approach to the use of relevance feedback as discussed, for example, by van Rijsbergen [19]. After each iteration of retrieval the user provides a relevance judgment for each document surrogate seen which permits calculation of a probability of relevance for each document property according to its distribution in relevant and non-relevant documents. In RMC's model this was part of Model 2, but in CEO Model 3 it will be a separate model. Rather than having the system mechanically update a user's subjective distribution, it is preferable to separate this analysis from Model 2. The mathematical details for the feedback expert have not yet been worked out, however, and so it will not be discussed further in this article or in part 2.

Finally it is necessary to consider the PIR system's distributions as upper level decision maker. In Lindley's CEO model the decision maker's distributions are normal, while the experts' distributions can take any form. The decision maker only uses measures of location and scale, such as mean and standard deviation, from each expert's distribution. Experts' means, being estimates of proportions, range from 0 to 1. For the decision maker's distribution for an expert's mean to be normal, however, the mean should range from $-\infty$ to $+\infty$. This problem can be solved by using the logistic transform of the expert's distributions, i.e., if the expert's mean is x, then the logistic transform is $\log x/(1-x)$. Once this transformation has been made it is reasonable to consider the decision maker's distribution normal in light of the assumption discussed above, i.e., that errors in probability assessment are normally distributed once transformed to log-odds.

The PIR system's prior, or initial, distribution fills an analogous role in CEO Model 3 to that of Model 0 in RMC's Model 3. As discussed above, for Model 0 there would often be little or no data. In this case Model 1 or Model 2 would have to be used alone in most formulations of Model 3 [1]. Bayesian CEO theory allows for the possibility of little or no prior data. The decision maker can have a vague, or noninformative, prior distribution, i.e., one which gives the same weight to all admissible values of the parameter. It this were a vague beta prior, then with each relevance judgment of the sort that would have supplied Model 0 data (see above), the prior distribution could be updated [47], thus ceasing to be noninformative. Combination of the models could always take place. Combination would be easier if the system's prior had a normal distribution, the natural conjugate prior distribution [50] for the system's normal likelihoods. A natural conjugate prior distribution for a likelihood function is one for which its associated revised, or posterior, distribution has the same distributional form as the prior, e.g., a beta distribution. Such distributions can be updated more easily than would be the case if the prior and posterior distributions had different distributional forms. A normal prior, however, could not be updated by relevance judgments. Using a beta prior with normal likelihoods would require numerical evaluation by the system, but this would present no difficulty.

An entirely different approach to providing the system with a prior distribution is suggested by Lenk and Floyd [51]. They propose constructing a multivariate normal prior based on a measure of association between vectors representing searches and documents. Such a prior distribution could be used with CEO Model 3.

9. THE PIR SYSTEM'S EVALUATION OF THE EXPERTS

RMC's Model 3 gives equal weight to Models 1 and 2 and considers the probabilities it works with to be calibrated. In a CEO model the decision maker must have an opinion of the relative expertise of the experts. In Lindley's model this is achieved through calibration. For example, if the decision maker thinks that an expert's mean tends to be too high, then the decision maker can center his or her distribution at a somewhat lower value. If the decision maker does not think the expert is very reliable, then he or she may increase the standard deviation of his or her distribution. A human decision maker could make

these calibration adjustments subjectively or through data; the PIR system will have to calibrate using data. Details will be given in part 2 (also see Thompson [26]). Briefly, indexers can be calibrated for each index term based on their past indexing performance with that term. This calibration at the level of individual terms allows Model 1 to be considered calibrated at the upper level as well, at least until it is recalibrated within the context of a particular search of several iterations. Users, on the other hand, will not, in general, use the system often enough for calibration data to be obtained, and so at the lower level Model 2 will be considered calibrated by default. At the upper level the PIR system will use an evaluation function to calibrate Models 1, 2, and 2*, the feedback expert.

The evaluation function will use the already calibrated Model 1 as a standard with which to compare the other models. It will assess the models' relative performance over all searches and earlier iterations of the current search using the ranking each model alone would have provided and relevance judgments. Even if Model 1, say, performed better than Model 2 overall, for some searches Model 2 might do better. If the evaluation function could detect this in early iterations, it could give Model 2 more weight in later iterations. The decision maker's variance for each model's mean will be increased or decreased according to how well the model did compared with Model 1. The higher the variance, the less impact the expert's mean will have. Just as Smeaton and van Rijsbergen [13] found a small sample problem due to the small number of relevant documents retrieved in an initial request, so here, too, there may be a similar problem when the evaluation function compares the models' performance within the context of a particular search, though presumably the improved retrieval resulting from the use of CEO Model 3 would provide more relevant documents in the initial iteration than was the case in their experiments.

10. INDEPENDENCE/DEPENDENCE IN CEO MODEL 3

Besides judging the experts' relative expertise, the PIR system must determine the dependence among the experts. At the upper level this is analogous to the problem in RMC's Model 3 of whether or not Models 1 and 2 can be assumed independent. The dependence problem has been discussed extensively in the recent CEO literature, e.g., Winkler [38], Lindley [45], Morris [42], Clemen and Winkler [52], but there has been no satisfactory resolution of the problem. Assuming independence combination would be straightforward, but frequently it cannot be assumed. Independence can be assumed for lower level CEO due to the use of term complexes, but at the upper level this particular solution is not available.

Huseby [53] suggests a promising approach to handling dependence which may be applicable to CEO Model 3. The dependence problem may be viewed as one of shared information. If two experts' distributions are based on completely disjoint sets of information, one would consider them to be independent. To the extent that the information overlapped, they would be dependent. In PIR this might work as follows. An indexer would have the whole document before him, but he would also be making probabilistic assessments on the basis of user-need properties. A user could introspect her information need, but would also make assessments on the basis of document properties. It would be impossible to exactly determine the amount of information overlap, but, based on empirical research, a rough estimate might be possible which might permit a more reasonable final probability than through assuming independence. However, assuming independence might be more reasonable than it first appears. Following Huseby's suggestion it can be argued, using RMC's model, that the effect of the shared information between Models 1 and 2 has already been incorporated in the Model 0 estimate. The information that an indexer and user share is the fact that the document being retrieved belongs to a class D of similar documents and that the user belongs to a class B of similar uses. This is precisely the information on which the Model 0 estimate is based [54].

11. CONCLUSION

In summary, PIR models developed within the machine intelligence paradigm suffer from problems of independence/dependence of term probabilities, small samples, and

validity of probabilities, and have usually been limited to consideration of only one kind of probabilistic evidence. RMC's Model 3 provides a framework in which to combine two earlier approaches to PIR, i.e., Models 1 and 2, but fails to resolve exactly how this combination is to be accomplished. At a purely formal level, i.e., without requiring the use of subjective probabilities, CEO Model 3 gives a solution to the problem of deriving Model 3. The feasibility of using subjective probabilities is undoubtedly controversial. While an examination of the psychological literature cannot decide the issue, it does suggest that human probability estimation can be accurate, that training can be effective, and that debiasing techniques may be helpful in improving estimates. It also shows that probability estimation is quite task dependent, implying that experiments will have to be done with probabilistic indexers and/or searchers before a final assessment of feasibility can be made.

If feasibility is demonstrated, subjective CEO Model 3, a machine-aided intelligence approach to PIR, would be able to confront the problems facing PIR models in general with certain advantages. The indexer or user could group dependent terms into complexes which were themselves approximately independent, thereby circumventing the independence/dependence problem at the level of individual terms. In assessing their subjective probabilities indexers or users would not be relying on samples, and so there would be no small sample problem, although the evaluation function might face a small sample problem when evaluating experts within the context of a particular search. User and indexer probabilities would have validity, since indexers and users would be competent to evaluate documents and information needs, respectively. Finally, through relevance judgments the PIR system would be able to calibrate and update indexers' initial probabilistic estimates.

Aside from the plausibility of using subjective probabilities, perhaps the major difficulty with CEO Model 3 is dependence among models. This problem also occurs with RMC's Model 3, or indeed any PIR model attempting to combine different sorts of evidence. More attention needs to be paid to this issue.

A final point worth noting is that although the PIR CEO model developed in this paper is based on RMC's Model 3, the CEO technique could also be used to combine evidence from experts other than those considered here. Belkin *et al.* [55], for example, describe a variety of possible experts that might provide evidence in information retrieval. Furthermore, although the use of distributions has been advocated in this paper, a PIR CEO model using probabilities would also be possible.

REFERENCES

1. Robertson, S.E.; Maron; M.E.; Cooper, W.S. Probability of relevance: A unification of two competing models for document retrieval. Information Technology: Research and Development, 1(1): 1-21; 1982.
2. Maron, M.E.; Kuhns, J.L. On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 7(3), 1960, 216-244.
3. Bookstein, A. Information retrieval: A sequential learning process. Journal of the American Society for Information Science, 34(5): 331-342; 1983.
4. Bookstein, A. Outline of a general probabilistic retrieval model. Journal of Documentation. 39(2): 63-72; 1983.
5. Cooper, W.S.; Huizinga, P. The Maximum Entropy Principle and its application to the design of probabilistic retrieval systems. Information Technology: Research and Development, 1(2): 99-112; 1982.
6. Cooper, W.S.; Maron, M.E. Foundations of probabilistic and utility-theoretic indexing. Journal of the ACM, 25(1): 67-80; 1978.
7. Croft, W.B.; Harper, D.J. Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 45(4): 285-295; 1979.
8. Harter, S.P. A probabilistic approach to automatic keyword indexing: Part 1. Journal of the American Society for Information Science, 26(4): 197-206; 1975.
9. Harter, S.P. A probabilistic approach to automatic keyword indexing: Part 2. Journal of the American Society for Information Science, 26(5): 280-289; 1975.
10. Radecki, T. Trends in research on information retrieval—The potential for improvements in conventional Boolean retrieval systems. Information Processing & Management, 24(3): 219-227; 1988.
11. Robertson, S.E.; Sparck Jones, K. Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3): 129-146; 1976.
12. Yu, C.T.; Salton, G. Precision weighting—an effective automatic indexing method. Journal of the ACM, 23(1): 76-88; 1976.
13. Smeaton, A.F.; van Rijsbergen, C.J. The retrieval effects of query expansion on a feedback document retrieval system. The Computer Journal, 25(3): 239-246; 1983.
14. Smith, L.C. Machine intelligence vs. machine-aided intelligence in information retrieval. In: Salton, G.;

Schneider, Hans-Jochen, editors. Lecture Notes in computer science: Research and development in information retrieval; Proceedings; Berlin, May 18-20, 1982. Berlin: Springer-Verlag; 1983: 263-274.
15. Wilson, P. Situational relevance. Information Storage and Retrieval, 9(7): 457-471; 1973.
16. Soergel, D. Is user satisfaction a hobgoblin? Journal of the American Society for Information Science, 27: 256-259; 1976.
17. Cooper, W.S. Indexing documents by Gedanken experimentation. Journal of the American Society for Information Science, 29(3): 107-119; 1978.
18. Bookstein, A. Relevance. Journal of the American Society for Information Science, 30(5): 269-273; 1979.
19. van Rijsbergen, C.J. Information retrieval. 2nd. ed. London: Buttersworth; 1979.
20. Kerlinger, F.N. Foundations of behavioral research. 2nd. ed. New York: Holt, Rinehart and Winston; 1973.
21. Yu. C.T.; Buckley, C.; Lam, K.; Salton, G. A generalized term dependence model in information retrieval. Information Technology: Research and Development, 2(4): 129-154; 1983.
22. Sparck Jones, K. Search term relevance weighting given little relevance information. Journal of Documentation, 35(1): 30-48; 1979.
23. Losee, R.M. Parameter estimation for probabilistic document retrieval models. Journal of the American Society for Information Science, 39(1): 8-16; 1988.
24. Spiegelhalter, D.J. A statistical view of uncertainty in expert systems. In: Gale, W.A., ed. Artificial intelligence and statistics. Reading, Mass.: Addison-Wesley; 1986; 17-55.
25. Tague, J.M. A Bayesian approach to interactive retrieval. Information Storage and Retrieval, 9: 129-142; 1973.
26. Thompson, P. Subjective probability, combination of expert opinion, and probabilistic approaches to information retrieval. Ph.D. dissertation. University of California, Berkeley; 1986.
27. Kahneman, D.; Slovic, P.; Tversky, A., editors. Judgment under uncertainty: Heuristics and biases. Cambridge, England: Cambridge University Press; 1982.
28. Hoffman, P.J. Cue-consistency and configurality in human judgment. In: Kleinmuntz, B., ed. Formal representations of human judgment. New York: Wiley; 1966.
29. Peterson, C.R.; Beach, L.R. Man as an intuitive statistician. Psychological Bulletin, 68(1): 29-46; 1967.
30. Gustafson, D.H. Evaluation of probabilistic information processing in medical decision making. Organizational Behavior and Human Performance, 4(1): 20-34; 1969.
31. Gustafson, D.H.; Kestly, J.J.; Ludke, R.L.; Larson, F. Probabilistic information processing: Implementation and evaluation of a semi-PIP diagnostic system. Computers and Biomedical Research, 6(4): 355-370; 1973.
32. Croft. W.B. Boolean queries and term dependencies in probabilistic retrieval models. Journal of the American Society for Information Science, 37(2): 71-77; 1986.
33. Bookstein, A. Implications of Boolean structure for probabilistic retrieval. Proceedings of the ACM Conference on Research and Development in Information Retrieval, 1985; 11-17.
34. Tague, J.; Nelson, M. Simulation of bibliographic retrieval databases using hyperterms. In: Salton, G.; Schneier, Hans-Jochen, eds. Lecture Notes in computer science: Research and development in information retrieval; Proceedings; Berlin, May 18-20, 1982. Berlin: Springer-Verlag, 1983: 194-208.
35. Losee, R.M.; Bookstein, A. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. Information Processing & Management, 24(3): 315-321; 1988.
36. Genest, C.; Zidek, J.V. Combining probability distributions: A critique and an annotated bibliography. Statistical Science, 1: 39-46; 1986.
37. Winkler, R.L. The consensus of subjective probability distributions. Management Science, 15(2): B-61-B-75; 1968.
38. Winkler, R.L. Combining probability distributions from dependent information sources. Management Science, 27(4): 479-488; 1981.
39. Winkler, R.L. Expert resolution. Management Science, 32(3): 298-303; 1986.
40. Morris, P.A. Decision analysis expert use. Management Science, 20(9): 1233-1241; 1974.
41. Morris, P.A. Combining expert judgments: A Bayesian approach. Management Science, 23(7): 679-693; 1977.
42. Morris, P.A. An axiomatic approach to expert resolution. Management Science, 29(1): 24-32; 1983.
43. Morris, P.A. Observations on expert aggregation. Management Science, 32(3): 321-328; 1986.
44. Lindley, D.V., Tversky, A.; Brown, R.V. On the reconciliation of probability assessments. Journal of the Royal Statistical Society, Series A, 142(2): 146-180; 1979.
45. Lindley, D.V. Reconciliation of probability distributions. Operations Research, 31(5): 866-880; 1983.
46. Lindley, D.V. Another look at an axiomatic approach to expert resolution. Management Science, 32(3): 303-306; 1986.
47. Bunn, D.W. Applied decision analysis. New York: McGraw-Hill; 1984.
48. Schaefer, R.E.; Borcherding, K. The assessment of subjective probability distributions: A training experiment. Acta Psychologica, 37: 117-129; 1973.
49. Gavaskar, U. A comparison of two elicitation methods for a prior distribution for a binomial parameter. Management Science, 34(6): 784-790; 1988.
50. Raiffa, H.A.; Schlaifer, R. Applied Statistical Decision Theory. Boston: Harvard Business School. Division of Research, 1961.
51. Lenk, P.J.; Floyd, B. Bayesian probabilities of relevance. Working Paper #85-97. New York: New York University. Graduate School of Business Administration, 1985.
52. Clemen, R.K.; Winkler, R.L. Combining economic forecasts. Journal of Business & Economics Statistics, 4(1): 39-46; 1986.
53. Huseby, A.B. The consensus problem. A retrospective approach. unpublished report of Center for Industrial Research, 1985.
54. Robertson, S.E. Personal communication, 1988.
55. Belkin, N.J.; Seeger, T.; Wersig, G. Distributed expert problem treatment as a model for information system analysis and design. Journal of Information Science, 5(5): 153-167; 1983.