When Is the Probability Ranking Principle Suboptimal?
Michael D Gordon; Peter Lenk
Journal of the American Society for Information Science (1986-1998); Jan 1992; 43, 1; ABI/INFORM Global
pg. 1

When Is the Probability Ranking Principle Suboptimal?

Michael D. Gordon and Peter Lenk
School of Business Administration, The University of Michigan, Ann Arbor, MI 48109-1234

The probability ranking principle retrieves documents in decreasing order of their predictive probabilities of relevance. Gordon and Lenk (1991) demonstrated that this principal is optimal within a signal detection—decision theory framework, and it maximizes the inquirer's expected utility for relevant documents. These results hold under three conditions: calibration, independent assessment of relevance by the inquirer, and certainty about the computed probabilities of relevance. We demonstrate that the probability ranking principle can be suboptimal with respect to expected utility when one of these conditions fails to hold.

Introduction

Probabilistic information retrieval (IR) systems compute a predictive probability that a given document will be relevant to a inquirer's information need. The subsequent selection of documents for retrieval is based on these probabilities. This article does not address the issue of computing or estimating predictive probabilities of relevance for an inquirer's request. Rather, we are concerned with retrieval policies after computing these probabilities. A retrieval policy is a rule to decide which documents to retrieve. The *standard retrieval policy* or, simply, the *standard policy,* orders the documents in a database by their predictive probabilities of relevance and retrieves documents according to this ranking. An important question is: In what sense and under what conditions is the standard policy optimal? Robertson (1977) showed that the standard policy has the largest expected number of relevant documents among all retrieval policies. However, the standard policy may not be optimal from a decision-utility theoretic perspective under all conditions.

Gordon and Lenk (1991) demonstrated that the standard policy has the greatest expected utility of all retrieval policies if three conditions hold. The first condition is that the system is *well calibrated.* Calibration ensures that the IR system's probabilities of relevance correspond to the inquirer's long-term assessment of relevance. If the system is well calibrated, then the inquirer can expect to find $p \times 100\%$ of the documents that have been assigned a predictive probability of p to be relevant. If the IR system is not well calibrated, i.e., if it is ill calibrated, then the probabilities that the system computes can be misleading and incorrectly rank documents. Trivially, an ill-calibrated system could be far from optimal because its computed predictive probabilities of relevance need not correspond to the long run performance of the IR system as judged by the inquirer.

The second condition is that the relevance of each retrieved document is independently appraised by the inquirer. This assumption implies that the inquirer's evaluation of one document will not change after evaluating another document. Additionally, if the system obtains the inquirer's evaluation, the predictive probability of relevance for other documents will not change in the light of this evaluation. Instead, the inquirer's query and the documents' descriptions are sufficient information for the system to produce predictive probabilities, and additional information, such as the inquirer's relevance judgements about other documents, is superfluous. Other authors, such as Bookstein (1983) and Lenk and Floyd (1988), have considered models with dependent assessments.

The third condition is that the probability of relevance is reported as a scalar by the system. Reporting a single number for a probability of relevance leads the inquirer to believe that the system knows these probabilities with certainty. In reality, the probabilities are estimated by the system rather than known with absolute certainty. Instead of a single number, the probability of relevance of a document for a query is better described by a probability distribution on the unit interval. This distribution can either be interpreted in the frequency sense as arising through sampling or in a subjective sense as describing the system's uncertainty

We dedicate this article to the memory of our mentor, colleague, and friend, Manfred Kochen. Address correspondence to the second author; both authors contributed equally to this article.

Received March 8, 1990; revised September 5, 1990; accepted October 30, 1990.

© 1992 by John Wiley & Sons, Inc.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

about the probability of relevance. If probability distributions, instead of scalars, are reported to the inquirer, it is not clear, however, how he or she should use these distributions to formulate retrieval policies. The retrieval policy may depend on parameters, such as the mean or median, that are calculated from these distributions. Different choices of parameters may result in different retrieval policies; thus, different documents would be retrieved.

In this article we examine the consequences of violating each of these three conditions. In particular, we present examples where the standard policy does not have the greatest expected utility when one of the conditions fails to hold.

This article has the following construction. The second section develops the background material needed for the examples that show if one of the above conditions is violated, the standard policy is suboptimal. This section formally defines terms and notation and introduces the critical concepts of calibration, refinement and Brier scores, which are used in evaluating forecasts, and also summarizes the optimality results of Gordon and Lenk (1991). The third section proposes methods to detect divergence from calibration. The fourth section analyzes the case where the system assumes independent appraisal of documents by the inquirer when, in fact, the inquirer's evaluations of documents have a dependency structure. In that circumstance, we show that the standard policy can be dominated by an alternative policy in terms of expected utility. The fifth section discusses the situation where the IR system's predictive probabilities are not known with certainty and are described by probability distributions. The final section contains concluding remarks.

This article critically examines possible difficulties with the standard policy, and the analysis applies to a variety of probabilistic learning and updating procedures, such as those described by Tague (1973), Bookstein (1983), Lenk and Floyd (1988), and Gordon (1991). The intent of this article is not to discredit probabilistic IR systems. On the contrary, we support their development. We believe that the standard policy is quite sensible and should not be discarded lightly.

Retrieval Policies

We first present background information about retrieval policies and formally define the terminology we use in this article. Then we describe the conditions when the standard retrieval policy is optimal in terms of expected utility.

Background

Probabilistic Information Retrieval Systems. Given an inquirer's query and a document, a probabilistic IR system computes the probability that the inquirer will find the document relevant. For document D_i let p_i be the predictive probability of relevance given the inquirer's query. We assume that for a given query, the IR system computes the predictive probability of relevance for each document in the database and that these probabilities are computed for single documents and not groups of documents. Thus, p_i is a function of the inquirer's query, the stored description of document D_i, and the IR system. It is the IR system's best guess about the relevance of D_i to the inquirer's information need as expressed by his or her query. We assume that these probabilities are in descending order

$$p_1 \geq p_2 \geq \cdots \geq p_M$$

for the M documents in the data base.

Let X_i indicate whether or not the inquirer finds D_i to be relevant to his or her query

$$X_i = \begin{cases} 1 & \text{if the inquirer finds } D_i \text{ relevant} \\ 0 & \text{if the inquirer finds } D_i \text{ nonrelevant} \end{cases} \quad (1)$$

Before the inquirer judges the documents, X_i is unknown.

Calibration, Refinement, and Brier Scores. Calibration is a long-term property of an IR system and is one indication of its performance. We will follow DeGroot and Fienberg's (1983) development of calibration. (Also, see DeGroot and Fienberg for earlier references.) Let $\xi(p)$ denote the conditional probability that the inquirer judges a document relevant given that the system reports a predictive probability of p

$$Pr(X_i = 1 \mid p) = \xi(p).$$

An IR system is *well calibrated* if $\xi(p) = p$ for all p, i.e., among all of the documents with predictive probability p, the proportion of relevant ones is p, for all p.

Calibration is needed for analyzing the properties of retrieval policies and for evaluating the performance of IR systems. However, it does not necessarily imply that the IR system is effective in helping the inquirer with his or her information needs: Two IR systems could be well calibrated, but one could be far superior to the other in delivering relevant documents. For instance, suppose the database has M items, and that, on average, r of the items are relevant to an inquirer. Then a system that reports the predictive probability $p = r/M$ for each of the items in the data base is well calibrated. However, such a system would not be useful to the inquirer in deciding which documents to retrieve. r/M would be an appropriate prior probability of relevance before the inquirer formulated a query to express his or her information need. After the system knows the inquirer's request, these prior probabilities should be updated by Bayes theorem. Thus, we see that, alone, calibration is insufficient to evaluate the performance of an IR system.

Calibration contributes to a scoring rule that can be used to measure the performance of an IR system.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Brier (1950) proposed a mean squared error rule for weather forecasting. The Brier score measures forecasting errors so that low Brier scores are desirable. Savage (1971) investigates conditions where the Brier score is appropriate and is a thoughtful article on rating forecasters. In our context, suppose that queries arrive sequentially. As in equation (1), let $X_{i,j}$ indicate whether or not the inquirer evaluates D_i relevant to the jth query, and let $p_{i,j}$ be the IR system's predictive probability for document D_i and query j. Suppose that there have been N queries to the system by an inquirer. Assume that the inquirer evaluates each of the M documents in the database for each query. If he or she only evaluates a subset of the documents, the definition of the Brier score has a simple modification, and the rest of the discussion is unchanged. The Brier score is a quadratic loss function between the inquirer's evaluation and the predictive probability supplied by the IR system

$$BS = (MN)^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N} (X_{i,j} - p_{i,j})^2. \qquad (2)$$

Next, suppose that the system computes only K *distinct* predictive probabilities, $p(1), p(2), \ldots, p(K)$. $p_{i,j}$, with subscripts, denotes the predictive probability of document D_i being relevant to query j, while $p(k)$ denotes the kth possible predictive probability. The $p_{i,j}$ are partitioned into classes defined by $p(k)$ according to the criterion $p_{i,j} = p(k)$. Let $n(k)$ be the number of times that the IR system has reported a predictive probability of $p(k)$ for the first N queries, let $r(k)$ be the number of relevant documents among these $n(k)$ documents. Define $f(k) = r(k)/n(k)$ to be the proportion of relevant documents given the system reports $p(k)$, and define $\nu(k) = n(k)/(MN)$ to be the proportion of forecasts that the system reports the prediction $p(k)$. Note that $n(k)$ and $r(k)$ depend implicitly on the number of queries, N. Because the $X_{i,j}$'s only take the values of zero or one, the Brier score can be written as

$$\begin{aligned} BS &= (MN)^{-1} \sum_{k=1}^{K} r(k)\{1 - p(k)\}^2 + \{n(k) - r(k)\}p(k)^2 \\ &= \sum_{k=1}^{K} \nu(k)\{f(k) - p(k)\}^2 + \sum_{k=1}^{K} \nu(k) f(k)\{1 - f(k)\} \\ &\equiv CS + RS \end{aligned}$$

where CS is the *calibration score,* and RS is the *refinement score.* The weights $\nu(k)$ in the calibration and refinement scores are the proportion of times that the system issues the forecast $p(k)$. Consequently, these scores are weighted averages of statistics that pertain to the individual $p(k)$'s.

The calibration score, CS, measures the discrepancy between the predictive probabilities, $p(k)$, and the observed relative frequencies, $f(k)$. If the inquirer independently assesses retrieved documents, then the empirical relative frequencies, $f(k)$, will tend to their true probabilities ξ as n increases, by the law of large numbers

$$\lim_{n(k)\to\infty} f(k) = \xi\{p(k)\}.$$

If the predictive probabilities are well calibrated, then the calibration score will tend towards zero as N becomes large. The larger the calibration score, the less reliably the system will rank the documents.

The refinement score measures the IR system's ability to correctly discriminate between relevant and nonrelevant documents by using the fixed set of predictive probabilities $\{p(k) : 1 \le k \le K\}$. Small values for the refinement score are desirable. The refinement score is a function of three components: the number of unique probabilities, K, that the IR system reports; the relative frequencies, $f(k)$'s, for those predictive probabilities; and the proportion, $\nu(k)$, that the system reports probability $p(k)$. The refinement score is maximized when $f(k) = 0.5$ for all k, which means that documents with any of the K predictive probabilities are equally likely to be relevant or nonrelevant. The refinement score is minimized when $f(k) = 0$ or 1 for all k. In a perfect, well-calibrated system, there are two predictive probabilities: $p(1) = 0$ and $p(2) = 1$. Then $f(1) = 0$ and $f(2) = 1$ in the long run, so the refinement score approaches zero. As a well-calibrated system uses more intermediate probabilities between 0 and 1, especially ones close to 0.5, the refinement score increases. For a fixed set of probabilities in a well-calibrated system, the refinement score is reduced if small and large probabilities are reported more frequently than intermediate probabilities. These frequencies of reporting are expressed by $\nu(.)$

DeGroot and Fienberg (1983) define refinement as: "[forecaster] A is at least as refined as [forecaster] B if we can artificially generate a well-calibrated forecaster with the same probability function [$\nu(k)$] as B simply by passing A's predictions through a noisy channel." DeGroot and Fienberg (1982) show that, for well-calibrated forecasters, if A is at least as refined as B, then the refinement score for A is less than or equal to that of B.

Retrieval Policies. We denote a retrieval policy by the indices of the documents, D_i, that it retrieves. Namely, a retrieval policy

$$R = \{r_1, r_2, \ldots, r_n\}$$

retrieves the n documents $D_{r_1}, D_{r_2}, \ldots, D_{r_n}$. The standard policy S retrieves the n documents which have the largest predictive probabilities of relevance. Since the probabilities are ordered,

$$S = \{1, 2, \ldots, n\}.$$

The number of relevant documents for policy R is

$$X(R) = \sum_{i \in R} X_i.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In the examples of the following sections, we use the statistical properties of retrieval policies. These properties follow from those of their constituent documents. Here, we present these properties. The mean of X_i (equation (1)) given p is $E(X_i|p) = \xi(p)$, and its variance is $\mathrm{Var}(X_i|p) = \xi(p)\{1 - \xi(p)\}$. The covariance between X_i and X_j given p_i and p_j is

$$\begin{aligned}\mathrm{Cov}(X_i, X_j | p_i, p_j) &= Pr(X_i = 1, X_j = 1 | p_i, p_j) - \xi(p_i)\xi(p_j) \qquad (3)\\ &\equiv \zeta_{i,j}\end{aligned}$$

We will use $\zeta_{i,i}$ and $\mathrm{Var}(X_i)$ interchangeably because variance is a special case of covariance: $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. If the inquirer's assessments of the documents' relevance are mutually independent, then

$$Pr(X_i = 1, X_j = 1 | p_i, p_j) = \xi(p_i)\xi(p_j)$$

and $\zeta_{i,j} = 0$ for $i \neq j$.

The expected number of relevant documents for policy R is

$$E\{X(R)\} = \sum_{i \in R} \xi(p_i),$$

and its variance is

$$\mathrm{Var}\{X(R)\} = \sum_{i \in R} \zeta_{i,i} + \sum_{i,j \in R; i \neq j} \zeta_{i,j}. \qquad (4)$$

If the inquire independently assesses all of the documents, then the $\zeta_{i,j}$ for $i \neq j$ are zero, which is one of the conditions for the standard policy to be optimal with respect to utility functions. Later in the article we will provide an example of dependent assessments where the standard policy is suboptimal.

All of our computations are conditional on a *single* inquirer. One reason for conditioning the analysis on individual inquirers is to avoid the suboptimal behavior that the standard policy can have when considering a population of inquirers, as shown by Cooper (1972) and illustrated in Gordon and Lenk (1991). Stirling (1977) further investigated these issues. In addition, utility functions are specific to individuals, not populations. However, it is important to note that two individuals with identical queries may have different information needs and may make different evaluations of the same documents.

Further, the computations of the means, variances and covariances are conditional on the IR system's, predictive probabilities $\{p_i\}$ for the documents $\{D_i\}$. In other situations, these parameters may be computed for populations of inquirers, for sets of queries, or without regard to the retrieval probabilities $\{p_i\}$. However, the meaning of the parameters in these situations is different than their meaning in the current context because of the different conditioning sets.

Optimality of the Standard Policy

For a well-calibrated system, the standard policy, S, maximizes the expected number of relevant documents because it selects the documents with the largest values of p_i. If the system is not well calibrated and if its predictive probabilities of relevance have the same ordering as those in a well-calibrated system, then the standard policy will still be optimal with respect to the expected number of relevant documents.

In addition to expected number, are there other criteria, such as expected utility, by which the standard policy is optimal? And are there situations where the standard policy should not be used? Bookstein (1977) used signal detection to show that a higher matching score, or retrieval status value, for a document does not necessarily imply that the document is more likely to be relevant than a document with a lower matching score. His results imply, in principle, that system designers should empirically validate the measurement scale underlying retrieval status values.

Gordon and Lenk (1991) demonstrated that, for well-calibrated probabilistic IR systems, predictive probabilities of relevance do follow the "bigger is better" rule. Consequently, the standard retrieval policy of selecting documents with the largest predictive probabilities is optimal from the signal detection point of view.

The signal detection model is a special case of utility theory. One of the implicit conditions of the signal detection model is that each relevant document has the same benefit regardless of the number of relevant documents that the inquirer has already obtained. This condition simplifies the analysis of retrieval policies, but it is not always appropriate. For example, the inquirer may obtain a "critical mass" of relevant documents, after which point additional relevant documents may be redundant. Or the cost of retrieval, in terms of time and intellectual energy, may increase as he or she examines more and more documents and finds fewer and fewer relevant ones.

Utility functions generalize the constant benefit or cost of the signal detection model. If the inquirer obtains diminishing marginal utility for additional documents, then it is possible that he or she prefers an alternative policy to the standard policy in terms of expected utility. This situation may arise if the inquirer is risk averse; in that case, he or she is willing to accept a policy that has a smaller expected number of relevant documents than the standard policy if the alternative policy is less variable in its results. The risk averse inquirer is willing to forgo some expected number of relevant documents for greater certainty about the number of relevant documents that he or she will retrieve.

For example, suppose that the inquirer would like to have about five relevant documents. Although having more than five documents would be nice, it is not of critical importance. However, having fewer than five documents is not acceptable. Suppose that the standard

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

policy has an expected value of 10 relevant documents, while an alternative policy, that retrieves the same number of documents as the standard policy, has an expected value of eight relevant documents. Further suppose that the alternative policy has a higher probability of retrieving five or more documents than the standard policy. For this inquirer's needs, the alternative policy might be preferred.

Gordon and Lenk (1991) showed that if the inquirer independently assesses the relevance of documents, if the system is well-calibrated, and if the predictive probabilities are reported as known scalars, then the standard policy has a greater expected utility than any other policy for all non-decreasing utility functions. Under these conditions the standard policy *stochastically dominates* any other policy. Stochastic dominance means that the standard policy S is more likely to provide more relevant documents than an alternative policy R. Formally, S stochastically dominates R if

$$Pr\{X(S) > x\} \geq Pr\{X(R) > x\} \quad \text{for all } x$$

where $X(S)$ and $X(R)$ are the number of relevant documents from the standard and alternative policies, respectively. Due to stochastic dominance, the example in the previous paragraph could never occur under the conditions of independence, calibration, and certainty. Nevertheless, these conditions do not necessarily hold. The remainder of the article investigates violations of these conditions.

Ill-Calibrated Systems

The first condition for the standard policy to be optimal with respect to expected utility is that the predictive probabilities are well calibrated. This section discusses the situation where the predictive probabilities are not well calibrated, i.e., they are ill calibrated. For ill-calibrated systems, the standard policy may not be optimal with respect to the expected number of relevant documents. If the conditional probabilities $\xi(p)$ are a monotonically increasing function of the predictive probabilities, p, then ordering documents based on p will be the same as ordering them based on $\xi(p)$. Therefore, the optimal policy based on the IR system's p will correspond to the optimal policy based on $\xi(p)$. On the other hand, if the ordering is not the same, then the standard policy could be much worse than an alternative policy. Recall that document D_1 is the document with the largest predictive probability, p_1. Suppose that the inquirer is actually less likely to find D_1 relevant than every other document in the database. Then the standard policy $S = (1, 2, \ldots, n)$ would be dominated with respect to expected value and utility by the alternative policy $R = (2, 3, \ldots, n + 1)$.

Detecting ill calibration requires the system to obtain feedback from the inquirer about his or her relevance assessments. A graph of the system's predictive probability, $p(k)$, versus $f(k)$, the relative proportion of documents that the inquirer judges relevant when the system assigns predictive probability $p(k)$, is a method of detecting ill calibration. If the system is well calibrated, then the graph should be a line through the origin with slope one. Deviations from the line indicate ill calibration. However, some deviations from the line would not change the optimality of the standard policy. If the $f(k)$'s are a monotonically increasing function of the $p(k)$'s, then the standard policy is optimal although the retrieval policy's statistics, such as its recall and precision, are not properly scaled to reflect the true performance of the policy. We note that the $f(k)$'s are sample estimates of $\xi\{p(k)\}$. Their sampling distributions must be considered when evaluating whether they significantly indicate departure from being well calibrated.

Dependent Relevance Assessments

The second condition for the optimality of the standard policy is that the inquirer's assessments of the documents are mutually independent. More often than not, the inquirer's evaluation of documents will have a dependency structure. For instance, consider a keyword based system where a set of identically described documents have many keywords in common with those of a query. The inquirer may judge all of the documents in the set to be relevant. Similarly, an identically described set of documents with few keywords in common with the query may all be judged to be nonrelevant by the inquirer. In both of these cases, the inquirer's relevance judgements are positively correlated for the documents within each of the sets. In this situation evaluating one document from the class is informative about the relevance of the other documents in the class. The opposite case may arise when two document subsets are never relevant to the same query. That is, if a document from one class is relevant, a document from the other class is nonrelevant, and conversely. In this case, the relevance judgements for documents from different classes are negatively correlated. This situation may be highly desirable because finding a nonrelevant document in one class implies that there is a large likelihood that the other class will have relevant documents. Less extreme cases have correlations between negative and positive one. In a different fashion, the presentation order of a set of documents may influence relevance judgements (Eisenberg & Barry, 1988).

Next, we present a simple example to illustrate that the standard policy can have a smaller expected utility than an alternative policy when there is a dependency structure in an inquirer's relevance judgements, even though the system is well calibrated.

Consider a well-calibrated system that has three documents D_1, D_2, and D_3 with predictive probabilities of relevance 0.5, 0.25, and 0.2, respectively. Consider two retrieval policies that retrieve two documents: the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

standard policy $S = \{1,2\}$ that retrieves D_1 and D_2, and the alternative policy $R = \{1,3\}$ that retrieves D_1 and D_3. As in equation (1), let X_i indicate whether or not the inquirer finds D_i to be relevant. Because the system is well calibrated

$$Pr(X_1 = 1) = 0.5, \quad Pr(X_2 = 1) = 0.25,$$
$$\text{and} \quad Pr(X_3 = 1) = 0.2.$$

The expected number of relevant documents, $X(S)$, from the standard policy is

$$E\{X(S)\} = 0.5 + 0.25 = 0.75,$$

and it is

$$E\{X(R)\} = 0.5 + 0.2 = 0.7$$

for the alternative policy. In terms of expected number of relevant documents, the standard policy is superior to the alternative.

However, suppose that the inquirer does not independently assess the relevance of the documents. In particular, suppose that the covariance between X_1 and X_2 is

$$\text{Cov}(X_1, X_2) = \zeta_{1,2} = 0.125,$$

and that the covariance between X_1 and X_3 is

$$\text{Cov}(X_1, X_3) = \zeta_{1,3} = -0.1.$$

Using equation (4), the variance in the number of relevant documents from the standard policy is

$$\text{Var}\{X(S)\} = p_1(1 - p_1) + p_2(1 - p_2) + 2\zeta_{1,2}$$
$$= 0.6875$$

while the variance for the alternative policy is

$$\text{Var}\{X(R)\} = p_1(1 - p_1) + p_3(1 - p_3) + 2\zeta_{1,3}$$
$$= 0.21.$$

Although the standard policy has a greater expected number of relevant documents, the inquirer is less certain of obtaining that number of relevant documents. A risk averse inquirer may prefer the alternative policy because he or she is willing to trade expected number for greater certainty in the number of relevant documents.

For this example, we next consider a specific utility function and demonstrate that the alternative policy has a greater expected utility than the standard policy. To begin, the probabilities and covariances of the three documents imply their joint distributions. For instance, using equation (3)

$$Pr(X_1 = 1, X_2 = 1) = \zeta_{1,2} + p_1p_2 = 0.125 + 0.125$$
$$= 0.25.$$

Also,

$$Pr(X_1 = 1, X_2 = 0) + Pr(X_1 = 1, X_2 = 1)$$
$$= Pr(X_1 = 1) = 0.5.$$

So

$$Pr(X_1 = 1, X_2 = 0) = 0.5 - 0.25 = 0.25$$

and so on.

Therefore, the joint probability distribution of X_1 and X_2 is

$$Pr(X_1 = 0, X_2 = 0) = 0.5 \quad Pr(X_1 = 0, X_2 = 1) = 0.0$$
$$Pr(X_1 = 1, X_2 = 0) = 0.25 \quad Pr(X_1 = 1, X_2 = 1) = 0.25.$$

The distribution of the number of relevant documents, $X(S)$, for the standard policy is

$$Pr\{X(S) = 0\} = 0.5; \quad Pr\{X(S) = 1\} = 0.25;$$
$$\text{and} \quad Pr\{X(S) = 2\} = 0.25.$$

Similarly, the joint probability distribution of X_1 and X_3 is

$$Pr(X_1 = 0, X_3 = 0) = 0.3 \quad Pr(X_1 = 0, X_3 = 1) = 0.2$$
$$Pr(X_1 = 1, X_3 = 0) = 0.5 \quad Pr(X_1 = 1, X_3 = 1) = 0.0.$$

The distribution of the number of relevant documents, $X(R)$, for the alternative policy is

$$Pr\{X(R) = 0\} = 0.3; \quad Pr\{X(R) = 1\} = 0.7;$$
$$\text{and} \quad Pr\{X(R) = 2\} = 0.0.$$

Next, consider utility functions of the form

$$U(x) = 1 - e^{-\delta x}$$

where δ is a positive number. This utility function is concave, passes through the origin, and approaches one as x tends towards infinity. The expected utility of relevant documents for the standard policy is

$$E\{U[X(S)]\} = 1 - (0.5 + 0.25e^{-\delta} + 0.25e^{-2\delta})$$
$$= 0.5 - 0.25e^{-\delta}(1 + e^{-\delta}).$$

The expected utility for the alternative policy is

$$E\{U[X(R)]\} = 1 - (0.3 + 0.7e^{-\delta}) = 0.7(1 - e^{-\delta}).$$

If $\delta = 2$, then $E\{U[X(S)]\} = 0.46$, and $E\{U[X(R)]\} = 0.61$, so that the alternative policy is preferred to the standard policy with respect to this utility function. Hence, the alternative policy is superior to the standard policy for this utility function, even though the standard policy has a larger expected value.

To summarize, when the three conditions detailed earlier are met, the standard policy stochastically dominates all other retrieval policies, and, thus, is superior to them on the basis of expected utility to an inquirer. These conditions are calibration, independent assessment of documents' relevance by the inquirer, and certainty in predictive probability. In this section, we have examined an example of violating the second condition and have seen that, in this case, an alternative retrieval policy is capable of providing an inquirer with greater expected utility than the standard policy.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Uncertainty in the Probability of Relevance

The third condition for the optimality of the standard policy is that the predictive probabilities of relevance are known with certainty. Generally, this condition is not realistic. Uncertainty can arise through the estimation of unknown parameters or through factors relating to the system's design, such as its query language, document representations, the appropriateness of its computational algorithms, and the various meanings of *relevance*. For instance, Robertson and Sparck Jones' (1976) relevance weighting of search terms, which is equivalent to van Rijsbergen's (1979) treatment, is based on frequency estimates. Clearly, these estimates will vary from sample to sample, and the inquirer should not be in the awkward position of having his or her retrieval being driven by these random fluctuations instead of substantive knowledge. In fact, in situations involving *many* uncertain variables, each of which must be combined in making the calculation, the certainty we place on a given predictive probability may be negligible.

Tague (1973) and Lenk and Floyd (1988) described the uncertainty in predictive probabilities of relevance by using probability distributions. The intent of these systems is to adequately model the uncertainty inherent in IR systems and to compensate for inadequate system design by incorporating the inquirer's judgements through Bayes theorem. Although these systems explicitly recognize the uncertainty in retrieval probabilities, they only communicate point estimates to inquirers without indicating the uncertainty of these point estimates. The point estimates are the means of the probability distributions.

Instead, IR systems should have a method of quantifying and communicating this uncertainty to the inquirer. Only reporting a point estimator, such as the mean, can be deceptive. In the following discussion, we will use an upper case P to indicate a random variable that has a distribution, while a lower case p refers to a scalar. For example, consider three predictive probabilities of relevance that are described by the following distributions:

(1) P_1 is 0.5 with probability 1

$$Pr(P_1 = 0.5) = 1.$$

(2) P_2 has the distribution

$$Pr(P_2 \leq x) = 3x^2 - 2x^3 \quad \text{for} \quad 0 \leq x \leq 1$$

with density

$$f_2(x) = 6x(1 - x) \quad \text{for} \quad 0 \leq x \leq 1.$$

(3) P_3 has a uniform distribution on the unit interval

$$Pr(P_3 \leq x) = x \quad \text{for} \quad 0 \leq x \leq 1$$

with density

$$f_3(x) = 1 \quad \text{for} \quad 0 \leq x \leq 1.$$

All three of these distributions have the same mean, $E(P_i) = 0.5$. However, they convey different information about the predictive probability of relevance. The distribution of P_1 is degenerate at 0.5, so the IR system is certain that P_1 is 0.5. The distribution of P_2 is symmetric about 0.5 with its mode at 0.5. Values close to 0.5 are more likely whereas values closer to zero or one are less likely. The distribution of P_3 gives intervals of equal length the same probability; the predictive probability is just as likely to be in the interval $(0, 0.2)$ as it is to be in $(0.5, 0.7)$ or $(0.8, 1)$. The distribution of P_3 expresses more uncertainty about the predictive probability than the other two distributions.

Suppose further that three documents, D_1, D_2, and D_3, have predictive probabilities P_1, P_2, and P_3, respectively. How should an inquirer use this information to retrieve one of the three documents? Different criteria affect the retrieval policy he or she will choose. If the inquirer finds an odds ratio of one for relevance to be attractive, he or she may prefer D_1 since it guarantees these odds. If he or she would rather have the document that is most likely to have a predictive probability between 0.6 and 0.8, then D_2 is preferred:

$$Pr(0.6 < P_1 < 0.8) = 0; \quad Pr(0.6 < P_2 < 0.8) = 0.248,$$
$$\text{and} \quad Pr(0.6 < P_3 < 0.8) = 0.2.$$

If he or she would rather have the document that is most likely to have a predictive probability greater than 0.8, then D_3 is preferred:

$$Pr(P_1 > 0.8) = 0; \quad Pr(P_2 > 0.8) = 0.104,$$
$$\text{and} \quad Pr(P_3 > 0.8) = 0.2.$$

If the inquirer prefers the document with the greatest expected predictive probability of relevance, he or she could randomly choose one of the documents since each distribution has the same expected value.

In summary, the preferred retrieval policy depends on the characteristic of the distribution that the inquirer uses in formulating the policy.

Next, we generalize the concept of Brier scores to include predictive probabilities of relevance described with uncertainty. This treatment allows IR systems that produce uncertain predictive probabilities to be evaluated. Then, we analyze examples of retrieval policies based on uncertain predictive probabilities. The examples highlight some further problems of reporting distributions instead of scalars. Following that, we propose using confidence intervals or highest predictive distribution intervals for communicating the uncertainty in a predictive probability. By examining these intervals, the inquirer can decide if the differences between such point estimators are sufficiently large to react differentially to the corresponding documents.

Brier Scores for Random Predictive Probabilities

Calibration does not easily generalize from scalar predictive probabilities to random predictive probabilities. If the system reports to the inquirer a point estimate derived from the distribution for the predictive probability, then the calibration of the point estimator

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

can be checked. If the entire distribution is presented to the inquirer, calibration cannot be checked unless the system knows how the inquirer is using the distribution to select documents.

However, the Brier score of equation (2), which is an overall measure of the performance of the IR system, does generalize to the case of uncertain predictive probabilities. It can be written in terms of three components. The first component measures the calibration of the *means* of the predictive distributions, the second component measures refinement, and the third component measures the total uncertainty or variance of the distributions. Consequently, for two system with identically calibrated means and refinements, the one with the greater certainty in predictive probabilities is preferred. As in the second section, let $X_{i,j}$ indicate whether or not document D_i is relevant to the inquirer's jth query, with one indicating relevance and zero indicating nonrelevance. Let $P_{i,j}$ be the predictive probability of relevance, and suppose that it has a distribution. For N queries and M documents the Brier score is the expected mean squared error between inquirer's relevance judgements, $X_{i,j}$, and the predictive probabilities, $P_{i,j}$:

$$BS = (MN)^{-1}\sum_{i=1}^{M}\sum_{j=1}^{N} E\{(X_{i,j} - P_{i,j})^2\}$$

$$= (MN)^{-1}\sum_{i=1}^{M}\sum_{j=1}^{N} E([\{X_{i,j} - E(P_{i,j})\} - \{P_{i,j} - E(P_{i,j})\}]^2)$$

$$= (MN)^{-1}\sum_{i=1}^{M}\sum_{j=1}^{N} \{X_{i,j} - E(P_{i,j})\}^2 + \mathrm{Var}(P_{i,j})$$

where the expectations are with respect to the distributions of the $P_{i,j}$'s. Because the Brier score is the expected mean squared error, small Brier scores are desirable.

Suppose that there are only K unique distributions, and let $P(k)$ be a random variable that has the kth distribution. As in the second section let $n(k)$ be the number of documents for which P_k was reported, let $r(k)$ be the number of those $n(k)$ documents that are relevant, let $f(k) = r(k)/n(k)$ be the proportion of relevant documents, and let $\nu(k) = n(k)/(MN)$ be the proportion of times that the system issues the prediction $P(k)$.

The Brier score can be written as

$$BS = (MN)^{-1}\sum_{k=1}^{K} r(k)[1 - E\{P(k)\}]^2 + \{n(k) - r(k)\}$$
$$\times [E\{P(k)\}]^2 + (MN)^{-1}\sum_{k=1}^{K} n(k)\mathrm{Var}\{P(k)\}$$
$$= \sum_{k=1}^{K} \nu(k)[f(k) - E\{P(k)\}]^2 + \sum_{k=1}^{K} \nu(k)f(k)$$
$$\times \{1 - f(k)\} + \sum_{k=1}^{K} \nu(k)\,\mathrm{Var}\{P(k)\}$$
$$\equiv CS + RS + \sum_{k=1}^{K} \nu(k)\,\mathrm{Var}\{P(k)\}.$$

The first term, CS, is the calibration score for the means of the $P(k)$'s. If the means are well calibrated and if the inquirer independently assesses the documents, then

$$\lim_{n(k)\to\infty} f(k) = E\{P(k)\}.$$

The second term, RS, is the refinement score. The third term penalizes the IR system for its uncertainty in predictive probabilities as measured by their variances. As the IR system's uncertainty about a predictive probability increases, its variance increases and inflates the Brier score. The weights, $\nu(k)$, are the proportion of time that the IR system presents the inquirer with the distribution for $P(k)$, so the third term is a weighted average of the variances for the individual $P(k)$'s and measures the total uncertainty, due to the distributions for the predictive probabilities, that the inquirer experiences over a number of queries.

Examples: Expectations, Likelihoods, and Utilities

If an IR system reports to the inquirer a distribution for the predictive probability of relevance, it is not clear how the inquirer should use it in formulating a retrieval policy. The inquirer could choose any of several parameters of the distribution to use in formulating a retrieval policy. For instance, if he or she is interested in minimizing the mean squared error loss between the point estimator and the predictive probability, then the mean of the distribution is a valid choice. Using the mean is also consistent with evaluating the system according to the Brier score. However, if mean absolute error is the criterion, then the median provides a better estimator of the predictive probability.

In this section we analyze retrieval policies when two different retrieval criteria are used: (1) retrieve the document with the largest expected predictive probability and (2) retrieve the document that is most likely to be relevant. These different criteria can lead to different orderings of documents for retrieval. We also relate these criteria to expected utilities.

Consider a system with two documents, D_1 and D_2, and corresponding probabilities of relevance, P_1 and P_2, which are random variables. We assume that the two random variables are independent. Suppose that the density of P_1 is

$$f_1(p) = 2p \quad \text{for } 0 \le p \le 1$$

with cumulative distribution function

$$F_1(p) = Pr(P_1 \le p) = \int_0^p f_1(u)\,du = p^2$$

For P_1, large predictive probabilities are more likely than small ones. Figure 1 graphs this density. The other elements of the graph are explained in the next section.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

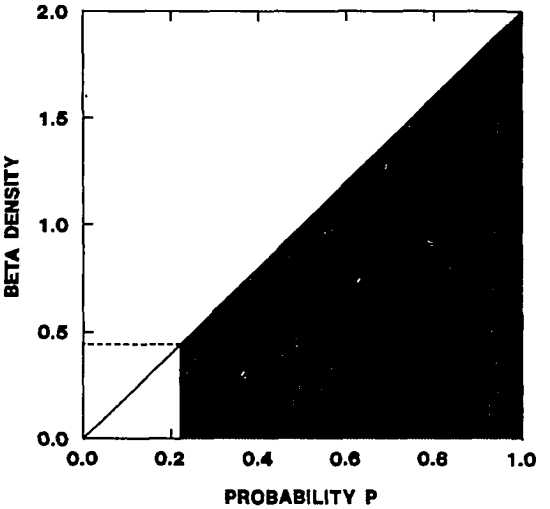

FIG. 1. Beta density for P_1 with 95% HPD interval.

The mean of P_1 is

$$E(P_1) = \int_0^1 p f_1(p)\,dp = 0.667.$$

Further, suppose that P_2 takes the values zero or one with probabilities

$$Pr(P_2 = 0) = 0.4 \quad \text{and} \quad Pr(P_2 = 1) = 0.6.$$

The mean of P_2 is 0.6, which is less than the mean of P_1, 0.667. Thus, document D_1 is preferred to document D_2 if retrieval is based on the expected values of the distributions for the predictive probabilities. However, the inquirer may prefer the document that is more likely to be relevant. In other words, he or she would prefer D_2 over D_1 if the probability of being relevant for D_2 is greater than that of D_1. In fact,

$$\begin{aligned} Pr(P_1 < P_2) &= Pr(P_1 < 0 \mid P_2 = 0)Pr(P_2 = 0) \\ &\quad + Pr(P_1 < 1 \mid P_2 = 1)Pr(P_2 = 1) \\ &= Pr(P_1 < 0) \times 0.4 + Pr(P_1 < 1) \times 0.6 \\ &= 0.6 \end{aligned}$$

(We used the independence of P_1 and P_2 in going from the second to the third line, and $Pr(P_1 < 0) = 0$ while $Pr(P_1 < 1) = 1$.) Because 0.6 is larger than 0.5, D_2 is more likely to be relevant than D_1. In fact, the odds ratio in favor of D_2 over D_1 is $1.5 = 0.6/0.4$, so that D_2 is 1.5 times more likely to be relevant than D_1. Therefore, if the inquirer chooses the document that is more likely to be relevant, then D_2 is preferred to D_1 even though D_1 has the greater expected probability of relevance.

These two retrieval criteria also have implications for expected utilities. Picking the document with the largest expected probability of relevance is equivalent to picking the document with the largest *unconditional* expected utility, as we now show. Define X_i as in equation (1). The *conditional* expected utility of selecting D_i given predictive probability P_i is

$$\begin{aligned} E\{U(X_i) \mid P_i = p\} &= U(0)(1 - p) + U(1)p \\ &= U(0) + \{U(1) - U(0)\}p \end{aligned}$$

The *unconditional* expected utility, $E\{U(X_i)\}$, integrates the conditional expected utility over the distribution for P_i:

$$\begin{aligned} E\{U(X_i)\} &= E[E\{U(X_i) \mid P_i\}] \\ &= U(0) + \{U(1) - U(0)\}E(P_i). \end{aligned}$$

Assuming that utility is an increasing function, we see that $E\{U(X_1)\} > E\{U(X_2)\}$ if and only if $E(P_1) > E(P_2)$.

The second retrieval criterion, selecting the document that is the most likely to be relevant, is equivalent to retrieving the document with the largest *conditional* expected utility. The conditional expected utility $E\{U(X_i) \mid P_i\}$ is a random variable because P_i is a random variable. It then makes sense to ask how likely it is that $E\{U(X_1) \mid P_1\}$ is less than $E\{U(X_2) \mid P_2\}$:

$$\begin{aligned} &Pr[E\{U(X_1) \mid P_1\} < E\{U(X_2) \mid P_2\}] \\ &\quad = Pr[U(0) + \{U(1) - U(0)\}P_1 \\ &\qquad < U(0) + \{U(1) - U(0)\}P_2] \\ &\quad = Pr(P_1 < P_2) \\ &\quad = 0.6 \end{aligned}$$

The last line follows from the previous calculation.

To summarize, with uncertain probabilities of relevance one document may be favored over another if the first's expected utility of relevance is greater than the second's. This criterion is equivalent to selecting the document with the greater expected probability of relevance. Alternatively, the second document may be favored over the first because it is more likely to have a greater conditional expected utility. This criterion is equivalent to selecting the document which is more likely to be relevant.

A second example uses the same distribution for P_1 and replaces the discrete distribution of P_2 with a continuous one. The distributions are from the family of beta distributions, which Tague (1973) used to describe the uncertainty in the predictive probabilities. This example demonstrates that the results of the first example do not require using discrete distributions.

The density function for the beta distribution with parameters α and β is

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1} \quad \text{for } 0 \le p \le 1$$

where α and β are positive numbers, and the gamma function is

$$\Gamma(r) = \int_0^\infty x^{r-1} \exp(-x)\,dx.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The first two moments of the beta distribution are

$$E[P] = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad E[P^2] = E[P] \times \frac{\alpha + 1}{\alpha + \beta + 1}.$$

We see that P_1, from the previous example, has a beta distribution with parameters $\alpha_1 = 2$ and $\beta_1 = 1$. Instead of the previous, discrete distribution for P_2, suppose that it has a beta distribution with parameters $\alpha_2 = 1/4$ and $\beta_2 = 1/6$, and that P_1 and P_2 are independent. With these values, the beta density of P_2 is "U" shaped with modes at zero and one. Figure 2 graphs this density. The other elements of the graph are explained in the next section. The moments of P_2 are

$$E[P_2] = 0.6 \quad \text{and} \quad E[P_2^2] = 0.5294.$$

As before, the mean predictive probabilities of relevance for D_1, 0.667, is greater than that for D_2, 0.6. However, D_2 is more likely to be relevant than D_1:

$$\begin{aligned} Pr(P_1 < P_2) &= \int_0^1 Pr(P_1 < p \mid P_2 = p) f_2(p)\, dp \\ &= \int_0^1 F_1(p) f_2(p)\, dp = \int_0^1 p^2 f_2(p)\, dp \\ &= E[P_2^2] = 0.5294 \end{aligned}$$

As in the first example, D_1 is preferred to D_2 if the inquirer uses the mean predictive probabilities (or the unconditional expected utility), while D_2 is preferred to D_1 if the inquirer would rather have the document that is more likely to be relevant (or more likely to have greater conditional expected utility).

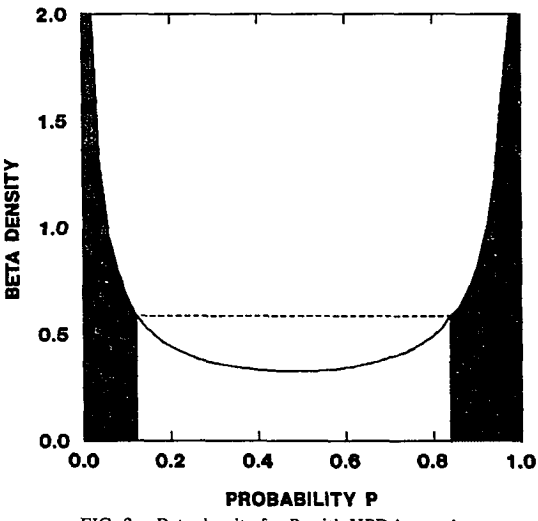

FIG. 2. Beta density for P_2 with HPD interval.

Confidence and HPD Intervals

Confidence intervals or highest predictive distribution (HPD) intervals (Berger, 1985, p. 140) can be used to communicate the variation in predictive probabilities of relevance. By so doing, we can reduce the possibility that the inquirer's decisions concerning the selection or ranking of documents are the result of statistical variation rather than substantive information.

A confidence interval is computed from the sampling distribution of estimated parameters, while an HPD interval is computed from the predictive distribution. In many settings, confidence and HPD intervals are mathematically equivalent, although their interpretations are different. The method for obtaining 95% confidence intervals ensures that 95% of all possible samples from the population will have intervals that contain the population parameter, while the subjective probability that the parameter is in the 95% HPD interval is 0.95, and this interval is as narrow as possible.

An HPD interval is computed from a density function by the following procedure. Draw a horizontal line across the graph of the density function. Orthogonally project the intersection points of the horizontal line and the density function onto the abscissa. The HPD interval is the region of the abscissa between the projected points where the horizontal line is below the density function. This HPD interval can, in fact, be the union of several, disjoint intervals. The content or level of this HPD interval is the area under the density and above this region. The horizontal line is adjusted until this area is equal to a specified value such as 0.95.

Figure 1 graphs the 95% HPD interval for the beta distribution with parameters $\alpha = 2$ and $\beta = 1$ and density

$$f(p) = 2p \quad \text{for } 0 \le p \le 1,$$

that is, the distribution of P_1 in the previous section. The shaded area has probability 0.95, and the interval on the P axis below the shaded region is the HPD interval. Since the density has a single mode, there is a single interval. Figure 2 graphs the beta density when $\alpha = 1/4$ and $\beta = 1/6$, which is the distribution for P_2 in the second example of the previous section. This density has two modes, so the HPD interval is the union of two intervals, one of which contains zero and the other contains one. The area above the HPD interval and under the density in Figure 2 is less than 0.95. However, by lowering the horizontal, dotted line the lengths of the two intervals increase, and the area under the density and over the intervals increases.

Confidence intervals also can be used when parameters are estimated. We illustrate their use with van Rijsbergen's (1979) model. In this model, a document has a binary description $D = \langle d_1, \ldots, d_K \rangle$. As customarily implemented, this vector includes only positions associated with terms used in the current query. Suppose that the IR system knows the probability that a keyword is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

present in the description of relevant documents and the probability that it is present in the description of nonrelevant documents:

$$\theta_k = Pr(d_k = 1 \mid \text{Rel}) \quad \text{and} \quad \phi_k = Pr(d_k = 1 \mid \text{Nonrel}).$$

The model assumes that the $\{d_k\}$ are independently distributed given relevance and nonrelevance.

The posterior odds ratio given document D is

$$\frac{Pr(\text{Rel} \mid D)}{Pr(\text{Nonrel} \mid D)} = \frac{Pr(\text{Rel}) \prod_{k=1}^{K} \theta_k^{d_k}(1 - \theta_k)^{1-d_k}}{Pr(\text{Nonrel}) \prod_{k=1}^{K} \phi_k^{d_k}(1 - \phi_k)^{1-d_k}}.$$

Van Rijsbergen (1979) defines the retrievel status value for document D as that part of the log posterior odds which depends on the description of D:

$$rsv(D) = \log\left(\frac{Pr(\text{Rel} \mid D)}{Pr(\text{Nonrel} \mid D)}\right) - \log\left(\frac{Pr(\text{Rel}) \prod_{k=1}^{K}(1 - \theta_k)}{Pr(\text{Nonrel}) \prod_{k=1}^{K}(1 - \phi_k)}\right)$$

$$= \sum_{k=1}^{K} d_k \left\{\log\left(\frac{\theta_k}{1 - \theta_k}\right) - \log\left(\frac{\phi_k}{1 - \phi_k}\right)\right\}.$$

The retrieval status value indicates the probability that D is relevant compared to other documents with different descriptions. Thus ordering documents based on their *rsv*'s corresponds to ordering documents based on $Pr(\text{Rel} \mid D)$.

In practice, however, θ_k and ϕ_k are both usually estimated from a sample, rather than known with certainty. As a consequence, we will obtain different values for $rsv(D)$ with different estimates of the θ_k's and ϕ_k's, and we should treat document rank orderings by *rsv*'s with some skepticism. Let n_k and m_k be the sample sizes for estimating θ_k and ϕ_k, respectively. (Losee (1988) and Fuhr and Hüther (1988) discuss estimation problems in IR.) Define $\hat{\theta}_k$ and $\hat{\phi}_k$ to be the maximum likelihood estimates of θ_k and ϕ_k: $\hat{\theta}_k$ is the proportion of times that keyword d_k appears in the random samples of relevant documents, and $\hat{\phi}_k$ is the proportion of times that keyword d_k appears in the random sample of nonrelevant documents. Since these estimators have sampling distributions, the estimated *rsv* also has a sampling distribution.

Next, we approximate the sampling distribution of *rsv*. The log-odds or logit function is

$$\text{logit}(p) = \log\left(\frac{p}{1 - p}\right).$$

The asymptotic distributions of $\text{logit}(\hat{\theta}_k)$ and $\text{logit}(\hat{\phi}_k)$ are obtained by a first-order Taylor series expansion about the population parameter. For simplicity, let $\hat{p}$ be the proportion of "successes" in n independent trial where the probability of success is p. Then the log-odds of $\hat{p}$ has asymptotic expansion

$$\text{logit}(\hat{p}) \approx \text{logit}(p) + \{np(1 - p)\}^{-0.5}\epsilon$$

where

$$\epsilon = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1 - p)}}.$$

By the central limit theorem, ϵ is asymptotically normal with mean 0 and variance 1.

Since the logit function is undefined when $\hat{p}$ is either 0 or 1, we consider the standard modification

$$\text{logit}'(\hat{p}, n) = \log\left(\frac{\hat{p} + \delta}{1 - \hat{p} + \delta}\right) \quad \text{where } \delta = n^{-1}.$$

Cox (1969, p. 33) analyzed this transform. The asymptotic mean of $\text{logit}'(\hat{p}, n)$ is $\text{logit}(p)$, and a nearly unbiased estimator of its variance is

$$S^2(\hat{p}, n) = \frac{(1 + \delta)(1 + 2\delta)}{n(\hat{p} + \delta)(1 - \hat{p} + \delta)}.$$

The estimated *rsv* measure is

$$\widehat{rsv}(D) = \sum_{k=1}^{K} d_k\{\text{logit}'(\hat{\theta}_k, n_k) - \text{logit}'(\hat{\phi}_k, m_k)\}.$$

Asymptotically, $\widehat{rsv}(D)$ is the sum of independent normal random variables. It is a consistent estimator of $rsv(D)$, and its asymptotic standard error is

$$SE\{\widehat{rsv}(D)\} = \left[\sum_{k=1}^{K} d_k\{S^2(\hat{\theta}_k, n_k) + S^2(\hat{\phi}_k, m_k)\}\right]^{0.5}$$

if the samples for estimating θ_k and ϕ_k are independent. The $(1 - \alpha) \times 100\%$ confidence interval for $rsv(D)$ is

$$\widehat{rsv}(D) \pm z_{\alpha/2}\, SE\{\widehat{rsv}(D)\}$$

where $z_{\alpha/2}$ is the upper $\alpha/2 \times 100$ percentile of the standard normal distribution.

Reporting the point estimator $\widehat{rsv}(D)$ along with its standard error $SE\{\widehat{rsv}(D)\}$ or its confidence interval conveys to the inquirer the uncertainty in the point estimator. Heuristically, if two confidence intervals overlap, then the inquirer may want to treat their *rsv*'s as being equivalent in formulating a retrieval policy. More rigorously, it is possible to compute the asymptotic distribution for the difference in the *rsv*'s for two documents. Suppose that document D_i is described by $D_i = \langle d_{i,1}, d_{i,2}, \ldots, d_{i,K}\rangle$. Further, suppose that the inquirer wants to determine if the *rsv*'s for documents D_i and D_j are statistically significantly different from each other. The point estimator for the difference in *rsv*'s is simply

$$\widehat{rsv}(D_i) - \widehat{rsv}(D_j) = \sum_{k=1}^{K}(d_{i,k} - d_{j,k})\{\text{logit}'(\hat{\theta}_k, n_k) - \text{logit}'(\hat{\phi}_k, m_k)\}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

with standard error

$$SE\{\widehat{rsv}(D_i) - \widehat{rsv}(D_j)\} = \left[\sum_{k=1}^{K} |d_{i,k} - d_{j,k}|^2 \{S^2(\hat{\theta}_k, n_k) + S^2(\hat{\phi}_k, m_k)\}\right]^{0.5}.$$

The $(1 - \alpha) \times 100\%$ confidence interval for $rsv(D_i) - rsv(D_j)$ is

$$\{\widehat{rsv}(D_i) - \widehat{rsv}(D_j)\} \pm z_{\alpha/2} SE\{\widehat{rsv}(D_i) - \widehat{rsv}(D_j)\}.$$

If this interval contains zero, then the inquirer could not reject the hypothesis that the *rsv*'s for D_i and D_j are statistically significantly different from each other. In this case, the inquirer may want to treat the two documents as having the same predictive probability of relevance in the retrieval policy.

More generally, suppose the inquirer is interested in an arbitrary linear combination of the *rsv*'s for I documents:

$$\psi = \sum_{i=1}^{I} c_i rsv(D_i).$$

Its point estimator is

$$\hat{\psi} = \sum_{i=1}^{I} c_i \widehat{rsv}(D_i) = \sum_{k=1}^{K} \left[\{\text{logit}'(\hat{\theta}_k, n_k) - \text{logit}'(\phi_k, m_k)\} \times \left(\sum_{i=1}^{I} c_i d_{i,k}\right)\right]$$

with standard error

$$SE(\hat{\psi}) = \left[\sum_{k=1}^{K} \left(\sum_{i=1}^{I} c_i d_{i,k}\right)^2 \{S^2(\hat{\theta}_k, n_k) + S^2(\hat{\phi}_k, m_k)\}\right]^{0.5}.$$

If the inquirer is concerned with more than one linear combination of the *rsv*'s, then he or she should use one of the methods of simultaneous inference. For instance, one of the more immediate methods of simultaneous inference is Bonferroni (Bickel & Doksum, 1977, p. 288). If there are u linear combinations of interest, then the total error α should be divided by u so that the overall confidence level for all u parameters is at least $1 - \alpha$. In other words, the simultaneous confidence intervals for the $\{\psi_v\}$ are

$$\hat{\psi}_v \pm z_{\alpha/2u} SE(\hat{\psi}) \quad \text{for } v = 1 \text{ to } u.$$

In practice, the linear combinations will probably be pairwise comparisons, and the number of linear combinations for I documents will be

$$u = \binom{I}{2}.$$

Discussion

The standard policy of retrieving first those documents with the highest predictive probabilities of relevance provides an important theoretical basis for information retrieval (Maron & Kuhns, 1960; van Rijsbergen, 1979). Robertson (1977) has shown that this policy yields the greatest expected number of relevant documents. Gordon and Lenk (1991) have shown that an inquirer's expected utility is also maximized by the standard policy, no matter what increasing utility function describes the inquirer. That is, the standard policy is still advisable even if the inquirer obtains diminishing value from each relevant document he or she retrieves from the collection.

In this article, we have analyzed the assumptions that underlie the optimality of the standard policy:

- Predictive probabilities are well calibrated.
- An inquirer independently assesses the relevance of the documents he or she retrieves.
- All predictive probabilities are reported with certainty.

Previous researchers have pointed to possible retrieval problems arising from the assumptions underlying a particular implementation of probabilistic IR systems. For instance, van Rijsbergen (1979) points out that calculating predictive probabilities of relevance is made easier, but less reliable, with the assumption that keywords are independently distributed given relevance or given nonrelevance of the document. Similarly, this calculation is shortened considerably by only considering keywords used in a query, rather than all keywords in the database as the formalism actually requires. Unlike these previous efforts that consider the assumptions that are applicable to a particular implementation or model of information retrieval, the assumptions that we analyze apply to all IR models.

The first assumption, that predictive probabilities are well calibrated, is most fundamental. Well calibration ensures that the system is predictively accurate. If the predictive probabilities are not well calibrated, that is, if they are ill calibrated, we cannot analyze a retrieval policy with complete assurance, and ill-calibrated predictive probabilities may not produce the optimal ordering of documents. In this case, it is helpful to know the degree of calibration of a system in order to determine the applicability of the predictions. The Brier score (Brier, 1950), which is used to assess the accuracy of forecasters, is applicable as well to IR. In brief, this score is the observed mean squared forecasting error and can be decomposed into two measures:

- Calibration, which measures the discrepancy between predictive probabilities and relative frequencies of successful predictions that correspond to these probabilities.
- Refinement, which measures the sensitivity of the system to discriminate between relevant and nonrelevant documents by using a fixed set of predictive probabilities of relevance.

Thus, the Brier score is a tool to assess the overall performance of the IR system. For instance, one use of the Brier score might be to determine the degree to which an assumption of independently distributed key-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

words, given relevance or given nonrelevance, reduces the accuracy of an IR system's predictive probabilities. Additionally, a Brier score may be an effective measure in choosing among alternative designs for a new system. For example, Brier scores may be used to compare two algorithms for calculating probabilities of relevance.

The second assumption necessary to ensure that the standard policy is optimal from the vantage of expected utility is that an inquirer independently assesses the relevance of each retrieved document. If not, the standard policy may have a smaller expected utility than an alternative policy. Our analysis evolves along lines similar to the decision-theoretic analysis of researchers such as Bookstein (1983), who showed that decision-theoretic costs can be reduced by incorporating the statistical dependencies among documents into a retrieval decision. We have extended this analysis to show that there are situations involving statistical dependencies in an inquirer's relevance judgments such that, even if the standard policy is advisable based on the expected cost of retrieval, it is inadvisable based on the expected utility to that inquirer.

The third assumption is that predictive probabilities are single numbers or scalars reported with certainty. Since many uncertain values enter into a prediction of a probability of relevance, we have argued that any predicted probability is most reliably described by a probability distribution. With these distributions, an inquirer or system that is formulating a retrieval policy can be better informed than by a single point estimator of this predictive probability.

However, new retrieval concerns arise when predictive probabilities are represented as distributions. For instance, one policy may be favored over a second policy on the basis of the expected number of relevant documents it retrieves even though the second policy is more likely to produce a greater number of relevant documents. Equivalently, the first policy will have a greater expected utility for the inquirer whereas the second will be more likely to have a greater conditional expected utility given the distributions for the predictive probabilities of relevance. As a result, there cannot be a single "right" way to use these distributions to formulate a retrieval policy. Similarly, different statistics derived from these distributions are preferred for different criteria for favoring a retrieval policy. For instance, basing a retrieval policy on a distribution's mean is consistent with minimizing mean squared error loss, while basing a retrieval policy on the median is consistent with minimizing absolute error loss.

If an IR system calculates a predictive distribution for a document and a query, we can use a generalization of the Brier score to assess its performance. We have extended the Brier scores so that it indicates:

- The overall, weighted variance associated with the predictive probability distributions computed by the system.
- The calibration of the means of these predictive distributions.
- The system's refinement score, which depends on the number of unique distributions reported by the system but not the distributions.

Thus, one IR system is preferred to another on the basis of lower overall variance, better calibrated means of the predictive distributions, or a better (lower) refinement score. Here, we are assuming "all other things being equal," e.g., better calibration assumes that one system is better than another with respect to calibration but that they have the same refinement and variance.

For IR systems that present predictive probabilities as a scalar value, we have recommended that these values be regarded as estimates of the true probability derived from an underlying sampling distribution. Thus, policies governing which documents to retrieve can be based on confidence intervals or Bayesian highest predictive intervals that are constructed around these point estimates. Presenting these intervals to an inquirer communicates their uncertainty and may suggest that two documents with different matching scores for the same query ought to be treated identically. Alternatively, we can compute the probability that these intervals for two documents overlap in order to decide whether we should regard one of these documents as being more likely than the other to be relevant.

The analysis in this paper demonstrates that the standard retrieval policy for probabilistic IR systems can often be suboptimal but does not suggest a viable alternative. Our intention is not to show that probabilistic IR systems are untenable, but rather to indicate their richness and subtlety. Kochen (1974, pp. 4–5) observed*

> [T]here were hopes that information theory would help us encode descriptors for data bases... that linear programming might help us design search strategies. This has not yet happened... It is possible that priorities of concern with file organization and data structures will be reversed toward more stress on what it takes to help the problem solver rather than how to optimally utilize computers and their programs.

We feel that the issues brought into focus by the analysis of probabilistic IR systems can assist in this reversal.

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer Verlag.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics.* San Francisco: Holden-Day.

Bookstein, A. (1977). When the most 'pertinent' document should not be retrieved—an analysis of the Swets' model. *Journal of the American Society for Information Science, 13,* 377–383.

*This quote was brought to our attention by a reviewer.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bookstein, A. (1983). Information retrieval: A sequential learning process. *Journal of the American Society for Information Science, 34,* 331–342.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Cooper, W.S. (1972). The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. Unpublished working paper, School of Librarianship, University of California.

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

DeGroot, M.H., & Fienberg, S.H. (1982). Assessing probability assessors: Calibration and refinement. In S.S. Gupta & J.O. Berger (Eds.), *Statistical decision theory and related topics III,* Vol. 1, (pp. 291–314). New York: Academic Press.

DeGroot, M.H., & Fienberg, S.H. (1983). The comparison and evaluation of forecasters. *The Statistician, 32,* 12–22.

Eisenberg, M., & Barry, C. (1988). Order effects: A study of the possible influence of presentation order on user judgements of document relevance. *Journal of the American Society for Information Science, 39,* 293–300.

Fuhr, N., & Hüther, H. (1989). Optimum probability estimation from empirical distributions. *Information Processing & Management, 25,* 493–507.

Gordon, M. (1991). Ranking large document collections by a state space search. *Information Processing and Management,* Vol. 27, No. 1, pp. 27–41.

Gordon, M., & Lenk, P. (1991). *A utility theoretic examination of the probability ranking principle in information retrieval. Journal of the American Society for Information Science, 42,* 703–714.

Kochen, M. (1974). *Principles of information retrieval.* Los Angeles: Melville Publishing.

Lenk, P., & Floyd, B. (1988). Dynamically updating relevance judgements in probabilistic information systems via users' feedback. *Management Science, 34,* 1450–1459.

Losee, R.M. (1988). Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science, 39,* 8–16.

Maron, M.F., & Kuhns, J.L. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of ACM, 7,* 216–244.

Savage, J.L. (1971). The elicitation of personal probabilities. *Journal of the American Statistical Association, 77,* 783–801.

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

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

Stirling, K.H. (1977). *The effect of document ranking on retrieval system performance: A search for an optimal ranking rule.* Ph.D. dissertation, University of California, Berkeley, May 1977. (Published as ERIC document IR 004 997, RIE Nov. 1977.)

Tague, J.M. (1973). A Bayesian approach to information retrieval. *Information Storage and Retrieval, 9,* 129–142.

Van Rijsbergen, C.J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.