A Utility Theoretic Examination of the Probability Ranking Principle in Information Retrieval

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

We challenge the probability ranking principle in information retrieval from the perspectives of (1) signal detection-decision theory and (2) utility theory. If three conditions are not met by an IR system that is producing predictive probabilities of relevance, then inquirers may incur costs that are too great by selecting first those documents that the system predicts have the highest probabilities of relevance. These three conditions are that predictive probabilities are well calibrated (predictively accurate); that they are reported with certainty; and that an inquirer independently assesses the relevance of all documents he or she retrieves. When these conditions are met, signal detection analysis with fixed decision-theoretic costs shows that the probability ranking principle is advisable. More generally, retrieval in adherence with the probability ranking principle is also advisable even when utility-theoretic costs (or benefits) that vary with the number of relevant documents retrieved are associated with retrieval. Specifically, we prove that the utility an inquirer receives from the relevant documents he or she retrieves is maximized by selecting those documents with the largest predictive probabilities of relevance.

Introduction

Information retrieval systems compute a matching function which assigns, to each document in a collection, a *retrieval status value.* Most IR systems then operate by furnishing an inquirer with those documents having the highest values.

There are situations, however, which suggest that it is not advisable for an IR system to follow this retrieval policy. First, a high retrieval status value may *not* necessarily suggest that a document is relevant. Instead, it may be necessary to know how retrieval status values are distributed within both the set of documents relevant to an inquirer's information need and within the set of nonrelevant documents to minimize expected retrieval costs. Second, for inquirers who want to avoid risk, alternative retrieval policies may be less risky (in terms of subjecting an inquirer to nonrelevant documents) even though, on average, they retrieve fewer relevant documents.

In this article we analyze probabilistic IR systems that

(1) assign *well-calibrated* predictive probabilities of relevance (predictively accurate retrieval status values) to documents
(2) report these probabilities with *certainty*
(3) assume an inquirer *independently assesses* the relevance of documents.

Probabilistic IR systems that retrieve documents with the highest predictive probabilities of relevance follow what we call the *standard (retrieval) policy.* For IR systems that have the above three properties, we prove the following:

(1) The standard policy is optimal in terms of minimizing decision theoretic retrieval costs, and we need *not* know the *distribution* of retrieval status values within the relevant and nonrelevant subsets of the collection.
(2) For all reasonable risk profiles—more specifically, for all utility functions that are nondecreasing with an increasing number of relevant documents—the expected utility an inquirer receives from following the standard policy exceeds the expected utility of any alternative retrieval policy.

We organize this article around these two findings. The next section presents background information. In the third section we show that signal detection theory—which can be used to show that higher retrieval status value are not always best from the vantage of minimizing retrieval costs—is in agreement with the standard policy applied to IR systems obeying our three assumptions. In the fourth section we examine utility theory, a more generalized cost-benefit structure that considers more than one document in assigning cost-benefits. We

Address correspondence to first author. Both authors contributed equally to this article.

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

show that the standard policy for IR systems built on the assumptions of calibration, certainty, and independent assessment is optimal from this perspective. In the fifth section we examine the assumptions underlying the class of IR systems we have analyzed. We discuss the necessity of these assumptions for reconciling retrieval based solely on probabilities of relevance with the counter-arguments of signal detection theory and utility theory. We also briefly discuss the limitations of IR systems that adhere to these assumptions.

Background

Probabilistic information retrieval systems have been widely discussed in the IR literature (see, for example, Bookstein (1985)). Although these systems may differ with respect to their underlying statistical assumptions and their methods of estimating parameters, they share the view that retrieval status values computed by a matching function have a probabilistic interpretation. That is, a retrieval status value with a probabilistic interpretation for document d_i indicates the *probability* that d_i is relevant to the inquirer's information need.* For emphasis, we indicate these retrieval status values by p, or p_i if we are specifically designating the ith document in the collection. Other IR systems compute matching functions based on similarity measures such as Jaccard's coefficient, Dice's coefficient, or the cosine coefficient to produce a measure of association between a document and a query. These measures do not have a probabilistic interpretation. For these, we use designations similar to rsv_i.

In both probabilistic and nonprobabilistic information retrieval, systems behave as desired when higher retrieval status values are associated with documents that are more likely to be relevant. Probabilistic and nonprobabilistic retrieval differ, however, in the way they compute these retrieval status values. Probabilistic systems formally incorporate information such as the probabilities of relevant and nonrelevant documents that contain a given keyword. When a keyword that a searcher uses in a query has a greater probability of occurrence in the relevant set of documents than in the nonrelevant set, the overall retrieval status value is adjusted in agreement with the laws of probability theory. On the other hand, nonprobabilistic retrieval systems do not explicitly obey these laws in calculating retrieval status values.

In this article, we analyze *retrieval policies.* A retrieval policy is a rule for determining which documents an inquirer should retrieve based on their retrieval status values. Further, a retrieval policy provides a rule for terminating the search based either on the total number of documents retrieved or the number of relevant documents retrieved. In the spirit of analyses by Cooper (1972), Robertson (1977), and Bookstein (1977), we examine the proper use of probabilistic retrieval status values and do *not* consider in detail methods to calculate them. We consider IR systems that assign a retrieval status value to individual documents, not sets of identically described documents. Further, we restrict our examination to systems that assign retrieval status values to all documents in a collection before any document is retrieved and that do not update these retrieval status values.

Intuitively, deciding which documents to retrieve for a particular query or rank ordering documents for presentation to the inquirer seems wholly dependent on *rsv*'s. Since a retrieval status value indicates the degree of association between a document representation and an inquirer's information need as represented by his or her query, the larger this value the better it is for the inquirer to retrieve the document. Despite this intuition, however, retrieving documents according to their retrieval status values is not necessarily justified. Arguments against retrieving solely on the magnitude of *rsv*'s can be based on signal detection theory or utility theory.

First, as Bookstein (1977) has shown using signal detection theory, too high a retrieval status value can actually be a strong indication that a document is *not* relevant. The same retrieval status value can be associated with either a relevant or with a nonrelevant document, so this value, alone, cannot indicate to the system that a document should be retrieved. Instead, the system must incorporate information concerning how likely it is for a relevant—and for a nonrelevant—document to have a given *rsv*. Analysis shows that even if, on average, nonrelevant documents have lower *rsv*'s than do relevant documents, a particularly high *rsv* may still be more likely to indicate a nonrelevant document if the *distribution* of *rsv*'s for nonrelevant documents is very spread out.

Second, utility theory considers situations in which individuals are *risk averse,* that is, situations in which gaining an additional dollar (or document) has less impact (in terms of change in magnitude of utility) than losing one. For instance, risk aversion explains why most of us would reject a $50,000 bet even if the odds are even or slightly in our favor: The consequences feel much more profound if we lose the bet than if we win it. Generally, in risk averse situations, one may have to choose between two alternatives, each of which is uncertain. The first alternative, on average, returns more money (or retrieves more relevant documents) than the second. But the distribution (in terms of money or number of relevant documents retrieved) is narrower for the second. Utility theoretic analysis shows that the better alternative for the risk averse user may be the one which, on average, returns less money (retrieves fewer relevant

*In some probabilistic IR systems, the calculation of the retrieval status value omits the prior probability that a document is relevant. In these situations, the system's calculated value for $Pr(\text{rel}\,|\,d_i)$ is easily recovered, at least in principle. Thus we include systems that make this type of calculation in our discussion of probabilistic information retrieval systems.

documents) but is less variable in the amount of money it returns (number of relevant documents it retrieves).

For the rest of this section, we describe different conditions that can underlie probabilistic information retrieval systems, and we review the most commonly discussed retrieval policy for these systems. We return to the problems posed by signal detection-decision theory and utility theory in the next two sections.

Calibration

Consider a probabilistic IR system that calculates, for each document d_i in a collection, a retrieval status value—i.e., predictive probability of relevance—p_i. For these probabilities to be considered *well-calibrated* (DeGroot & Fienberg, 1983), the probabilities that the system reports must correspond to the long-term proportion of relevant documents that would be retrieved when retrieving all documents with the same calculated value. That is, the probabilities predicted by the system correspond exactly to the inquirers' behavior in the long run. For example, on 70% of the occasions that the system reports a predictive probability of relevance of 0.70, the inquirer will actually find the associated document relevant. Similarly, if the system reports a predictive probability of relevance of one, then the user will, in fact, find that document relevant. More generally, the user will find p * 100% of the documents to be relevant over all instances when the system calculates a predictive probability of relevance of p. Thus, a well-calibrated system makes predictions that accurately reflect an inquirer's judgments over a period of time. A "perfect" well-calibrated system would correctly report only probabilities of zero or one. Thus, such a system makes accurate calculations on every occasion. Such a system would require that documents and information needs be completely and perfectly represented and that the matching function be capable of making complete use of this information. Such systems remain an ideal.

Certainty

We assume as well that the IR system reports predictive probabilities of relevance with certainty. That is, no range of values or distribution of values is associated with p. For example, the system may report that $p = 0.25$ but would not simply report that p lies between 0.18 and 0.29 or that p has any value in this range with identical probability.

Independent Assessment

We assume that an inquirer's relevance judgment for one document never affects the relevance judgment made for another document. This means that the predictive probability of relevance assigned to a document is based solely on its characteristics, the characteristics of the query, and never information pertaining to what other documents have been judged relevant already. The IR systems we discuss, then, need only consider a single document at a time in assigning predictive probabilities of relevance.

Standard Retrieval Policy

To maximize the expected number of relevant documents in a set of n retrieved documents, a well-calibrated IR system in which an inquirer independently assesses the relevance of the documents he or she retrieves should follow what we deem the *standard (retrieval) policy:* Select documents strictly according to their predictive probabilities of relevance, those with higher values of p_i first.

Robertson (1977) defended the optimality of the standard policy under certain assumptions in discussing the probability ranking principle. For well-calibrated IR systems that assign predictive probabilities of relevance independently to documents, the optimality of picking exactly n documents by the standard policy is explained by the expected number of relevant documents the policy retrieves. In particular, suppose n documents are retrieved according to some retrieval policy and their predictive probabilities of relevance are $p_1, p_2, \ldots, p_n$, respectively. Since the system is well-calibrated, the expected number of relevant documents that will be retrieved is $\Sigma_{i=1}^{n} p_i$. Therefore, the expected level of recall is $\Sigma_{i=1}^{n} p_i/N$, where N is the number of relevant documents in the collection. Similarly, $\Sigma_{i=1}^{n} p_i/n$ is the expected level of precision. The optimality of the standard policy follows immediately from the fact that this policy selects the n documents with the largest p_i's. As a result, no other set of documents can produce a larger value for the sum of the p_i's.

Notice that a system that is *not* well-calibrated but rank orders documents identically to a system that is well-calibrated will be just as effective in terms of expected recall and precision as the standard policy. For instance, if a well-calibrated system reports predictive probability of relevance p_i for each document, d_i, in the system, then a system that reported predictive probability $p_i/2$ would identically rank order documents even though each predictive probability would be only one half as large as it should be. In this event, the n documents with the largest predictive probabilities of relevance would still produce the largest sum, $\Sigma_{i=1}^{n} p_i$, even though this sum is no longer properly scaled to produce accurate estimates of recall or precision. Of course, a system that is not well calibrated is not guaranteed to correctly rank order documents by their actual probabilities of relevance. Thus, a well-calibrated system guarantees that the standard policy will retrieve, on average, the maximum number of relevant documents, though any other system that reports predictive probabilities of relevance that are monotonic with these well-calibrated probabilities will be just as effective.

Cooper (1972) has shown that adherence to the probabilistic ranking principle can be problematic if the probabilities used are aggregated measures from a user population and the stopping rule for retrieval is that a certain number of relevant documents has been retrieved. Since Cooper's paper was not published, we provide an example similar to those he analyzed. Suppose that there is a class, C, of inquirers each of whom issues the identical query equally often. But, for two-thirds of this class, i.e., subclass C_1, documents d_1, d_2, and d_3 are relevant and for the other third, subclass C_2, document d_4 and d_5 are relevant. Then, the probabilities of relevance that should be assigned are: two-thirds for d_1 through d_3; and one-third for d_4 and d_5. The standard policy, then, would dictate that the first five documents to be retrieved should be d_1, d_2, d_3, d_4, and d_5, and in that order (with ties in predictive probabilities arbitrarily broken).

If a retrieval policy is based on stopping when a specified number of relevant documents has been retrieved, the standard policy can be worse, even on average, than alternative policies. For instance, using the standard policy, an inquirer from C_1 desiring exactly one relevant document would have a search length (number of nonrelevant documents retrieved (Cooper, 1968)) of zero, and an inquirer from C_2 would have a search length of three. Thus, by the relative proportion of C_1 and C_2, the *expected* search length for C is one. Alternatively, with a retrieval order of d_1, d_4, d_2, d_3, d_5 the expected search length is only one-third. Similar difficulties occur in a variety of more complex retrieval situations. Consequently, when probabilities are aggregated across classes of inquirers, the standard policy may be suboptimal.

We assume for the remainder of this article that an IR system produces well-calibrated probabilities calculated with respect to an individual inquirer, reports these probabilities with certainty, and users evaluate the relevance of documents independently of each other. Under these assumptions, we show that the standard policy is optimal from the vantage of both signal detection theory and utility theory. In the two sections that follow, we consider each of these perspectives. We then return to a discussion of our assumptions in the conclusion.

Signal Detection Theory and Decision Theory

A challenge to the advisability of the standard policy arises when signal detection theory coupled with decision theory is used for analysis. Bookstein (1977) explored IR system performance from a decision theoretic vantage by using Swets' (1963) signal detection model. He showed that higher retrieval status values are not necessarily better. We critically examine probabilistic retrieval in light of Bookstein's analysis. We follow Bookstein's development in our discussion.

There are four different outcomes when one can either retrieve or not retrieve a document and that document may be relevant or may be nonrelevant. Following decision theory, and remaining completely general, each of these outcomes may be assigned a (possibly) different nonnegative cost. Table 1 summarizes this situation.

This cost structure can be used in either of two ways: for *ranking* documents by expected costs or for *classifying* them. The first way, we calculate, for every document, its expected *difference* in cost from being retrieved versus not being retrieved and rank order documents on this basis, lower expected difference first. An inquirer would then retrieve from the top of this list until he or she is satisfied or decides to terminate the search. The second way, a cutoff or threshold is used to divide the ranked list into two portions—a portion containing just those documents at or below threshold, and a portion containing the remaining documents. In particular, if the threshold is zero, the former list contains all and only those documents whose expected cost when the document is retrieved is less than its expected cost when it is not. In other words, the IR system is

TABLE 1. Costs associated with retrieving and not retrieving documents.

	Relevant	Nonrelevant
Retrieve	C_{11}	C_{12}
Do Not Retrieve	C_{21}	C_{22}

classifying these documents as those that should be retrieved. The discussion that follows assumes than an IR system is making such classifications.

With the signal detection theory model, no matter what retrieval status value, rsv_i, an IR system produces for document d_i, it cannot use this information alone to determine for certain whether to retrieve d_i. Instead, it also must know the distribution of retrieval status values in both the sets of relevant and nonrelevant documents. Without information about these distributions, it is impossible to tell whether a particular retrieval status value indicates a relevant document or whether nonrelevant documents are more likely to have this retrieval status value. Next we use signal detection theory to analyze IR systems. To begin with, our analysis applies to both probabilistic and nonprobabilistic IR systems, and then it focuses on well-calibrated probabilistic systems.

Following Bookstein's discussion, we assume that $f_R(rsv)$ is the probability density of a relevant document that has retrieval status value rsv. That is, it is the density of the "signal" having retrieval status value rsv (Fig. 1). Recognize that $f_R(rsv)$ tells us, after the retrieval status values have been calculated for all the relevant documents in a collection, how these retrieval status values are distributed. Thus, document and query characteristics are reflected by this density. $f_{NR}(rsv)$ defines the corresponding density function with respect to the set of nonrelevant documents. That is, it is the density for the "noise." (Here and later, we indicate by R the event that a document is relevant and by NR the event that it is not.)

Consider a document, d_i, that has retrieval status value rsv_i for some query. The system must decide whether to retrieve d_i for the user. By retrieving d_i, the expected cost to the user will be

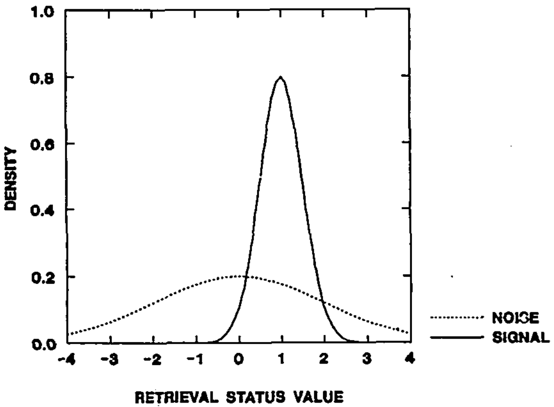

FIG. 1. Signal and noise densities.

$$C_{11} * Pr(R|rsv_i) + C_{12} * Pr(NR|rsv_i).$$

Similarly, the expected cost of not retrieving this document is

$$C_{21} * Pr(R|rsv_i) + C_{22} * Pr(NR|rsv_i).$$

Decision theory advises that d_i be retrieved if and only if the expected cost of retrieving d_i is less than the expected cost of not retrieving it. That is, the document with retrieval status value rsv_i should be retrieved if and only if

$$C_{11} * Pr(R|rsv_i) + C_{12} * Pr(NR|rsv_i) < C_{21} * Pr(R|rsv_i) + C_{22} * Pr(NR|rsv_i).$$

Algebraic manipulation and an application of Bayes' rule give an equivalent decision rule: Retrieve d_i if and only if

$$\frac{Pr(R)}{Pr(NR)} * \frac{f_R(rsv_i)}{f_{NR}(rsv_i)} > \frac{C_{12}\text{-}C_{22}}{C_{21}\text{-}C_{11}}. \quad (1)$$

What is interesting about decisions made according to this rule is that, in some circumstance, (1) the retrieval status values for the documents that should be retrieved are lower than for those that should not; and (2) that there may even be two noncontiguous ranges of retrieval status values whose associated documents should be retrieved.

An interpretation of equation (1) helps explain these seeming anomalies. The left-hand side of this equation is the product of two terms. The left-more term, $Pr(R)/Pr(NR)$, gives the prior odds of a document being relevant to the given query. This term is the same for any document and conveys how likely an arbitrary document is to be relevant. The right-more term, $f_R(rsv_i)/f_{NR}(rsv_i)$, the likelihood ratio, changes for different retrieval status values, rsv_i. Thus, the likelihood ratio determines the magnitude of the left hand side of equation (1). Specifically, when the likelihood ratio is greater than one, the left-hand side of equation (1) has a value greater than the prior odds ratio, indicating that such a document is more likely than an arbitrary document to be relevant. Such documents are the best candidates for retrieval.

Geometrically, we see that the likelihood ratio is greater than one everywhere the density for the "signal" (i.e., relevant documents) exceeds the density for the "noise" (i.e., nonrelevant documents) (Fig. 1). The greatest values of equation (1) are restricted to these regions, and, thus, may occur for retrieval status values other than the highest. For instance, the density associated with a retrieval status value of 2.5 is substantially greater for the noise than for the signal, indicating that a document with such a retrieval status value is very likely to be nonrelevant. To be precise, it is the relative magnitude of the variances of $f_R(\cdot)$ and $f_{NR}(\cdot)$, together with a document's retrieval status value, that determines if that document should be retrieved.

This situation can be likened to tuning a radio in order to listen to a certain radio station. The frequency scale is not constructed so that "bigger is better," so it does no good to tune the radio to too high a frequency because then the signal broadcast by the station one wants to hear will be dominated by "noise" from higher frequencies. Instead, one must tune as closely as possible to the frequency of the station one wants to hear.

Overall, then, signal detection theory together with decision theory suggest that we *cannot* take on face value retrieval policies of the following form to maximize expected recall from the first n documents retrieved: With cosine retrieval, retrieve the n documents with greatest cosine; with simple matching, retrieve the n documents with greatest simple matching score; etc. (Such rules are analogous to always attempting to get the best signal from a preferred radio station by tuning the receiver to the far right side of the dial.)

Instead, the *distribution* of these matching scores with respect to both the set of relevant and nonrelevant documents must be considered in theoretically defending a retrieval policy. Thus, the signal detection theory model is using these distributions to empirically validate that its measurement scale is used to correctly classify documents. Other IR models make no use of these distributions. Implicitly, these models assume without empirical validation that retrieval status values are ordinally scaled; that is, they assume that a document with a higher *rsv* should always be retrieved before one with a lower *rsv*.

To sum up to this point, signal detection-decision theoretic analysis dictates that the documents one should retrieve not necessarily have the greatest retrieval status values. On the other hand, if a cutoff is used to classify documents with the standard policy, the documents the system will retrieve will be exactly those documents with the greatest retrieval status value (i.e., with predictive probability of relevance greater than the cutoff). In other words, the signal detection-decision theory result suggests the possibility of rejecting the standard policy.

However, if probabilistic retrieval is based on well-calibrated probabilities, the standard retrieval policy is consistent with signal detection-decision theoretic analysis. We show this next.

Theorem. Define $C = (C_{12} - C_{22})/(C_{21} - C_{11})$ and assume a probabilistic system is well-calibrated. Then, consistent with the standard policy, just those documents whose retrieval status value, p_i, is greater than the cutoff $C/(1 + C)$ should be retrieved.

Proof. We can rewrite equation (1) to give the conditions when a document with retrieval status value rsv should be retrieved. This condition is

$$\frac{Pr(R)}{Pr(NR)} * \frac{f_R(rsv_i)}{f_{NR}(rsv_i)} > C.$$

Equivalently,

$$\frac{f_R(rsv_i)}{f_{NR}(rsv_i)} > \frac{C}{\dfrac{Pr(R)}{Pr(NR)}}.$$

Since

$$f_R(rsv_i) = \frac{Pr(R\,|\,rsv_i) * Pr(rsv_i)}{Pr(R)},$$

and

$$f_{NR}(rsv_i) = \frac{Pr(NR\,|\,rsv_i) * Pr(rsv_i)}{Pr(NR)},$$

the condition for retrieval may be written as

$$\frac{Pr(R\,|\,rsv_i)}{Pr(NR\,|\,rsv_i)} * \frac{Pr(NR)}{Pr(R)} > C * \frac{Pr(NR)}{Pr(R)}.$$

By cancelling the common factor, we obtain the condition

$$\frac{Pr(R\,|\,rsv_i)}{Pr(NR\,|\,rsv_i)} > C.$$

For probabilistic IR systems that are well calibrated, a retrieval status value reflects the long-term probability that a document with such a retrieval status value is relevant. That is,

$$Pr(R\,|\,p_i) = p_i \quad \text{and} \quad Pr(NR\,|\,p_i) = 1 - p_i.$$

(Keeping with our convention, we indicate retrieval status values for probabilistic IR systems similarly to p, or p_i if we are referring to a particular document.) Therefore, our condition for retrieval becomes

$$\frac{p_i}{1 - p_i} > C.$$

Equivalently, retrieval is advisable whenever*

$$p_i > \frac{C}{1 + C}.$$

Thus, retrieval status values that are produced by well-calibrated probabilistic IR systems can be used to classify documents for retrieval with respect to the cutoff $C/(1 + C)$. Those documents whose retrieval status values are above the cutoff are to be retrieved, those below are not. This is precisely the way in which documents are classified for retrieval by the standard retrieval policy. Unlike the case for other similarity measures, any document, d_i, for which $rsv_i = p_i > C/(1 + C)$ will have an expected cost that is lower when d_i is retrieved than when it is not, thus justifying its retrieval. Thus, the standard policy is brought into

*This equivalence depends on the fact that $C > 0$. This condition holds when the cost of a false identification is greater than the cost of a true identification, whether that identification was that a document was relevant or nonrelevant.

agreement with the analysis for classifying documents for retrieval by signal detection and decision theory.

In much the same way, if a document listing that is rank ordered by decreasing (*non*probabilistic) retrieval status values is provided to the inquirer to allow him or her to select as many documents as he or she wishes to see, the ordering is not necessarily the best with respect to maximizing recall. The order will be the best, however, if the retrieval status values are well-calibrated probabilities and the inquirer independently assesses documents.

Utility Theory

Utility theory is a generalization of the decision theoretic cost-benefit structure we considered in the last section.† Whereas decision theoretic costs remain the same for each document considered for retrieval, the utility theoretic benefit of each relevant document retrieved varies according to the number of previously retrieved relevant documents. With this generalization to allow changes in the cost-benefit structure over time comes another challenge to the wisdom of guiding retrieval by the standard policy.

In this section, we study this challenge. First, we introduce utility functions in relation to a familiar (monetary) context. Next, we discuss how theory indicates that maximizing the expected number of dollars (documents) one obtains in an uncertain situation does not guarantee that one's utility will also be maximized. Then, we prove the main result of this section: that the standard retrieval policy is optimal with respect to any nondecreasing utility function, given the system is well-calibrated, reports probabilities with certainty, and an inquirer independently assesses the relevance of every document he or she retrieves.

A utility function is a mapping from one scale measuring units of value to another. For instance, a utility function might map dollars (a real measuring scale) to "utils" (a fictitious one). One desires to increase his or her utility, whereas it is desirable to hold down one's decision-theoretic costs. Also unlike the costs associated with decision theory, the additional utility of the next additional dollar is not constant. For instance, the mapping $U(x) = \log(x)$ indicates that the (total) utility of x dollars grows logarithmically, not linearly, as one would expect if every additional dollar added equal value to total utility. In the same way, each additional relevant document that an IR system retrieves does not necessarily provide the inquirer with the same additional utility.

Of particular interest are individuals who are *risk averse.* For risk averse individuals, the additional utility one receives decreases with every additional dollar. In other words, $U'(k)$—the *marginal utility* added by the kth dollar—is decreasing. Many individuals display risk averse behavior (Kahneman, Slovic & Tversky, 1982).

Being risk averse may mean preferring a lesser expected amount of money to a greater expected amount that is associated with greater risk. For instance, a risk averse individual would prefer a fixed sum of money to a gamble yielding an uncertain sum with the same expected value. Before proving our main result, we present three examples which illustrate risk averse utility functions. These examples are increasingly sophisticated. Each can be ignored without a loss of continuity.

Example 1.

Consider a user with utility function $U(x) = \log(x)$ who has the choice between (1) \$5 with certainty and (2) any (integer) dollar amount between \$4 to \$6 with equal probability. The expected utility of the certain payout is $U(5) = \log(5) = 1.606$. The expected utility of the uncertain payout is $E(U(X))$ where X is a uniformly distributed random variable taking integer values from \$4 to \$6.

$$\begin{aligned} E(U(X)) &= \sum_{x=4}^{6} \log(x) * p(X = x) \\ &= \sum_{x=4}^{6} \log(x) * (1/3) \\ &= 1.595. \end{aligned}$$

Since $1.595 < 1.606$, this individual would rationally prefer the certain \$5 to the uncertain payout.

Risk aversion applies also to situations in which one is faced with two uncertain situations, instead of a single uncertain situation to be considered against a certain alternative. Consider the following example.

Example 2.

An individual receives utility $U(x) = 1 - e^{-ax}$ from x dollars, where a is some positive number. The concave plot of this utility function (i.e., the negative second derivative of $U(X)$) reveals the user is risk averse with respect to any amount of money.* Suppose further that he or she has the choice of two amounts of money, X_1 or X_2, each to be determined by a different random process. In particular, X_1 will be selected arbitrarily from one normal distribution and X_2 from another. Suppose the expected value returned to X_1 is greater than the expected value returned by X_2. Still, since the user is risk averse, it may turn out that he or

†Decision theory associates costs with actions, and utility theory associates benefits. Thus, one desires low decision-theoretic costs but high utilities.

*A utility function may be concave over some subset of its domain and convex over others. Thus, strictly speaking, one must qualify a statement about risk aversion as applying to the appropriate subset of the domain of the utility function.

she receives greater utility, on average, from X_2 than X_1 if X_1 is more spread out.

Let X_1 and X_2 be defined by the following normal distributions

$$X_1 \sim N(\mu_1, \sigma_1^2)$$
$$X_2 \sim N(\mu_2, \sigma_2^2)$$

That is, X_1 and X_2 have mean μ_1 and μ_2 and standard deviation σ_1 and σ_2, respectively. Let μ_1 be greater than μ_2. Then, $E(U(X_2)) > E(U(X_1))$ whenever $\sigma_1^2 - \sigma_2^2 > 2/a * (\mu_1 - \mu_2)$. That is,

$$\sigma_1^2 - \sigma_2^2 > 2/a * (\mu_1 - \mu_2) \quad \text{iff}$$
$$a^2/2 * (\sigma_1^2 - \sigma_2^2) > a * (\mu_1 - \mu_2) \quad \text{iff}$$
$$-a\mu_1 + a^2/2 * \sigma_1^2 > a\mu_2 + a^2/2 * \sigma_2^2 \quad \text{iff}$$
$$\exp(-a\mu_1 + a^2/2 * \sigma_1^2) > \exp(-a\mu_2 + a^2/2 * \sigma_2^2) \quad \text{iff}$$
$$1 - \exp(a\mu_1 + a^2/2 * \sigma_1^2) < 1 - \exp(-a\mu_2 + a^2/2 * \sigma_2^2).$$

But

$$1 - \exp(-a\mu_1 + a^2/2 * \sigma_1^2)$$

is the expected value of $U(X_1)$. Similarly,

$$1 - \exp(-a\mu_2 + a^2/2 * \sigma_2^2)$$

is the expected value of $U(X_2)$. So,

$$E(U(X_2)) > E(U(X_1)).$$

Thus, we have shown that, for a risk averse user with a particular utility function, a gamble that, on average, returns *less* money than another has *greater* average utility if its variance is also significantly less (more exactly, if it is less by at least $2/a * (\mu_1 - \mu_2)$ where μ_i is the expected value of gamble X_i and a is a parameter that defines the individual's utility function).

Example 3.

More generally, let a gamble have an uncertain payout of X, with X having expected value μ and variance σ^2. For an individual with wealth W, where W is a great deal larger than X, we can use a Taylor series to approximate the expected utility of the wealth of this individual after taking the gamble:

$$U(W + X) \approx U(W + \mu) + U'(W + \mu)(X - \mu) + \frac{1}{2} U''(W + \mu)(X - \mu)^2.$$

Taking expectations, we obtain

$$E(U(W + X)) \approx E(U(W + \mu)) + E[U'(W + \mu) \times (X - \mu)] + E\left[\frac{1}{2}U''(W + \mu)(X - \mu)^2\right].$$

Or,

$$E(U(W + X)) \approx U(W + \mu) + U'(W + \mu)E(X - \mu) + \frac{1}{2}U''(W + \mu)E((X - \mu)^2).$$

Equivalently,

$$E(U(W + X)) \approx U(W + \mu) + \frac{1}{2}U''(W + \mu)\sigma^2.$$

Thus, when U″(·) is negative, the expected utility of one's wealth after taking one gamble may exceed his or her expected utility after taking another gamble that has a larger expected payout but also a larger variance.

To summarize, our examples of utilities have portrayed situations in which an individual is to receive an uncertain sum of money but may prefer either the same sum, if that sum is guaranteed, or another uncertain sum that has a smaller expected value. In addition, the concept of utility applies to IR systems. We consider the situation in which it is uncertain how many relevant documents an inquirer will retrieve and he or she is risk averse with respect to retrieving relevant documents. That is, the additional utility to the inquirer of an additional relevant document is outweighed by the loss in utility by retrieving one fewer relevant document. Such an inquirer acts conservatively, just as does an individual who is risk averse with respect to money.

Intuitively, an *inquirer* can be risk averse for different reasons. For instance, he or she may already have a "critical mass" of relevant documents and may find additional ones less useful, on average, than those already retrieved. (This corresponds to a decrease in C_{21}, the cost of not retrieving a relevant document, with the retrieval of additional relevant documents (Table 1).) In another case, the additional effort, measured by number of nonrelevant documents retrieved, may increase with each relevant document retrieved. (This corresponds to an increasing values for C_{11}, the cost of retrieving a relevant document.) The marginal utility of the next relevant document is thus reduced because of this greater effort that must be expended.

As we have stated, a utility function (for relevant documents) captures information similar to the costs associated with decision theory. However, utilities change with respect to the number of relevant documents retrieved whereas decision theoretic costs are constant for each document. In other words, a linear utility function is equivalent to a set of constant decision-theoretic costs. Further, as we have seen, the standard retrieval policy is optimal with respect to the expected number of documents retrieved and with respect to constant decision-theoretic costs. But, we have seen with our monetary example that incorporating *changing* costs (specifically, a nonlinear utility function) can cause a gamble with a smaller expected payout to be preferred to another gamble with a higher expected payout. Similarly, if a retrieval policy is based on expected utility, rather than constant costs, we might suspect that it is not always optimal to follow the standard retrieval policy. For instance, for a risk averse inquirer and two different retrieval policies used to select 100 documents, one that retrieves 30 relevant documents, on average, and the other 40 on average, the first policy still might be preferred if it has a much tighter distribution in the number of relevant documents it retrieves. Next, we show that such situations *cannot* arise.

Specifically, we will prove that if a probabilistic IR system is well-calibrated and reports these probabilities with certainty, and if an inquirer independently assesses the relevance of retrieved documents, then the standard retrieval policy is optimal with respect to expected utility for *all* nondecreasing utility functions. Our proof follows from the fact that the standard policy *stochastically dominates* all other retrieval policies. Intuitively, if one retrieval policy stochastically dominates another, then, for any number of relevant documents, the former policy is at least as likely to retrieve *more* than that number of relevant documents than is the latter.

More formally, consider two retrieval policies, $policy_1$ and $policy_2$, each operating on the same document database for the same query. Suppose each has been used, independently, to retrieve m documents. Let X_i be a random variable indicating the number of relevant documents retrieved by $policy_i$. Then,

$$F_{i,m}(x) = Pr(X_i \leq x \mid m \text{ documents retrieved})$$

is the *cumulative distribution* of X_i. That is, $F_{i,m}(x)$ indicates the probability that the total number of relevant documents retrieved by $policy_i$ after m documents have been retrieved is at most x.

If $F_{1,m}(x) \leq F_{2,m}(x)$ for all x, we say that the distribution of relevant documents for $policy_1$ *stochastically dominates* the distribution for $policy_2$. Intuitively, this means that $policy_1$ is more likely than $policy_2$ to have retrieved a large number of relevant documents. For instance, if $m = 20$, then $F_{1,m}(19) \leq F_{2,m}(19)$ means that $policy_1$ is more likely than $policy_2$ to have retrieved all 20 relevant documents; $F_{1,m}(18) \leq F_{2,m}(18)$ means that $policy_1$ is more likely than $policy_2$ to have retrieved at least 19 relevant documents; etc.

Thus, a strong condition for preferring $policy_1$ to $policy_2$ is that $policy_1$ stochastically dominates $policy_2$.* Moreover, stochastic dominance provides a means for evaluating retrieval policies in terms of their expected utilities, as we prove next. Finally, as we prove after that, stochastic dominance corresponds to the notion that the standard policy is "the best" retrieval policy.

*This is a strong condition since we may also prefer $policy_1$ to $policy_2$, in our example, even if $F_{1,m}(x) \leq F_{2,m}(x)$ only holds when x is at least 2. This condition has the interpretation that, if one wants more than one relevant document, $policy_2$ is preferred; but if one wants more than two, $policy_1$ is preferred.

Theorem. Let $U(x)$ be an arbitrary, nondecreasing utility function, where x indicates a number of relevant documents. Let X_i be a random variable indicating the number of relevant document retrieved by $policy_i$ after n documents have been retrieved. If X_1 stochastically dominates X_2, then the expected utility of following $policy_1$, $E(U(X_1))$, is greater than or equal to $E(U(X_2))$, the expected utility of following $policy_2$.

Proof. Assume X_1 and X_2 are both discrete random variables taking values between 0 and n. If X_1 and X_2 have the same distribution, then $E(U(X_1)) = E(U(X_2))$, as we need to show. So, assume the distributions of X_1 and X_2 are not the same. Since X_1 stochastically dominates X_2, there is some m, $0 \leq m \leq n$, for which $F_1(m)$ is strictly less than $F_2(m)$.

We wish to show that:

$$E(U(X_1)) = \sum_{i=0}^{n} Pr(X_1 = i) * U(i)$$

$$= \sum_{i=0}^{m-1} Pr(X_1 = i) * U(i) + \sum_{i=m}^{n} Pr(X_1 = i) * U(i)$$

$$\geq E(U(X_2)) = \sum_{i=1}^{m} Pr(X_2 = i) * U(i)$$

$$= \sum_{i=0}^{m-1} Pr(X_2 = i) * U(i) + \sum_{i=m}^{n} Pr(X_2 = i) * U(i).$$

Equivalently, we wish to show

$$\sum_{i=m}^{n} [Pr(X_1 = i) - Pr(X_2 = i)] * U(i) \geq \sum_{i=0}^{m-1} [Pr(X_2 = i) - Pr(X_1 = i)] * U(i).$$

But,

$$\sum_{i=0}^{m-1} Pr(X_1 = i) + \sum_{i=m}^{n} Pr(X_1 = i) = 1 = \sum_{i=0}^{m-1} Pr(X_2 = i) + \sum_{i=m}^{n} Pr(X_2 = i),$$

so the probabilities—i.e., the bracketed terms—on both sides of the ">" in the last inequality sum to the same amount. Therefore, since $U(X)$ is nondecreasing,

$$\sum_{i=m}^{n} [Pr(X_1 = i) - Pr(X_2 = i)] * U(i)$$

$$\geq U(m) * \sum_{i=m}^{n} [Pr(X_1 = i) - Pr(X_2 = i)]$$

$$= U(m) * \sum_{i=0}^{m-1} [Pr(X_2 = i) - Pr(X_1 = i)]$$

$$\geq \sum_{i=0}^{m-1} [Pr(X_2 = i) - Pr(X_1 = i)] * U(i),$$

as we wished to show.*

*For those familiar with Lesbesgue integration, a more direct proof is

$$E[U(X_i)] = \int_{-\infty}^{\infty} U(x)\,dF_i(x) = \int_{0}^{1} U[F_i^{-1}(p)]\,dp$$

by the change of variables, $p = F_i(x)$, and letting $F_i^{-1}(p) = \inf\{x \mid F_i(x) \geq p\}$. But, $F_2(x) \geq F_1(x)$ iff $F_1^{-1}(p) \geq F_2^{-1}(p)$. Thus, since $F_2(x) \geq F_1(x)$ and $U(\cdot)$ is nondecreasing, $E(U(X_1)) \geq E(U(X_2))$.

To summarize, we have just seen that if retrieval policy$_1$ stochastically dominates retrieval policy$_2$, the expected utility to a user of following the first policy ensures that his or her expected utility will be greater than for policy$_2$, no matter what the precise form of his or her (nondecreasing) utility function.

We show next that the standard retrieval policy stochastically dominates all other retrieval policies. Thus, it should always be preferred on the basis of expected utility.

Theorem. Assume that policy$_1$ selects n documents in a sequence, sequence$_1$ = $\langle d_{11}, d_{12}, d_{13}, \ldots, d_{1n}\rangle$, where each d_{1i} is some document in the collection. Similarly, suppose that policy$_2$ selects, independently of policy$_1$, n documents according to sequence$_2$ = $\langle d_{21}, d_{22}, d_{23}, \ldots, d_{2n}\rangle$, where d_{2i} is also a document from the same collection, possibly one of the documents d_{11} through d_{1n}. Assume the probability that the jth document in sequence$_1$ is relevant is $p_{1,j}$, the probability of relevance for the jth document in sequence$_2$ is $p_{2,j}$, and that these probabilities are well-calibrated, reported with certainty, and the inquirer independently assesses the relevance of every document he or she retrieves. It follows that, if $p_{1,j} \geq p_{2,j}$ for $1 \leq j \leq n$, then X_1, the distribution of relevant documents for policy$_1$, stochastically dominates X_2, the distribution for policy$_2$.

Proof. We show by induction that $F_{1,n}(\cdot) \leq F_{2,n}(\cdot)$ for all n.

Basis: By assumption

$$p_{1,1} \geq p_{2,1}.$$

Since these probabilities are well calibrated,

$$F_{1,1}(0) = 1 - p_{1,1} \leq F_{2,1}(0) = 1 - p_{2,1}$$

and

$$F_{1,1}(1) = 1 = F_{2,1}(1).$$

Thus, $F_{1,1}(\cdot) \leq F_{2,n}(\cdot)$.

Induction: Suppose

$$F_{1,j}(\cdot) \leq F_{2,j}(\cdot) \quad \text{for } j \leq m.$$

We must show that the inequality also holds for $j = m + 1$. That is, we must show that

$$Pr(X_1 \leq k \mid m + 1 \text{ documents retrieved}) \leq Pr(X_2 \leq k \mid m + 1 \text{ documents retrieved}).$$

(Hereafter, we abbreviate our notation so that a conditioning event is written similarly to "$m + 1$," instead of "$m + 1$ documents retrieved.") Recall that, by definition, $p_{i,m+1}$ = the probability that the m + 1st document retrieved by policy$_i$ is relevant, and such probabilities are well-calibrated. Then, by the definition of $F_{i,m}(x)$,

$$Pr(X_1 \leq k \mid m + 1) \leq Pr(X_2 \leq k \mid m + 1) \quad \text{iff}$$

$$F_{1,m}(k - 1) + Pr(X_1 = k \mid m) * (1 - P_{1,m+1}) \leq F_{2,m}(k - 1) + Pr(X_2 = k \mid m) * (1 - P_{2,m+1}).$$

(This inequality uses the facts that (1) the event $\{X_{m+1} \leq k\}$ is the union of two events:

$\{X_m < k\} U \{X_m = k$ and the m + 1st document retrieved is not relevant$\}$ and (2) the inquirer independently assesses the relevance of every retrieved document.) Using

$$Pr(X_i = k \mid m) = F_{i,m}(k) - F_{i,m}(k - 1),$$

we can rewrite this equation and distribute factors to obtain an equivalent inequality

$$F_{1,m}(k - 1) * p_{1,m+1} + F_{1,m}(k) * (1 - p_{1,m+1}) \leq F_{2,m}(k-1) * p_{2,m+1} + F_{2,m}(k) * (1 - p_{2,m+1}). \quad (2)$$

We note that the right-hand side of this inequality decreases in value with increasing $p_{2,m+1}$ because $F_{2,m}(k-1) \leq F_{2,m}(k)$. $p_{2,m+1}$ cannot exceed $p_{1,m+1}$, however, since this would be in violation of the constraint we have placed on sequence$_2$ that $p_{i,j} \geq p_{2,j}$, for all j. Letting

$$p_{2,m+1} = p_{1,m+1} = p,$$

we see that

$$F_{1,m}(k - 1) * p + F_{1,m}(k) * (1-p) \leq F_{2,m}(k-1) * p + F_{2,m}(k) * (1 - p),$$

since, by the inductive hypothesis, $F_{1,m}(\cdot) \leq F_{2,m}(\cdot)$. Thus, inequality (2) holds when we have selected a value of $p_{2,m+1}$ that minimizes its right-hand side, and so it holds for all other values, too. Thus,

$$F_{1,m+1}(\cdot) \leq F_{2,m+1}(\cdot),$$

as we wished to show.

It follows immediately that the standard policy—in which the document with highest possibility of relevance is retrieved first, the document with next highest probability is retrieved second, etc.—stochastically dominates any alternative policy that retrieves documents whose probabilities are ordered by some other decreasing sequence of probabilities. This dominance is a result of the fact that, for the standard policy, the probability of relevance of the ith document retrieved is at least as great as the probability of the ith document retrieved by the alternative policy, since both policies retrieve documents by some decreasing order of their probabilities of relevance.* Note that an alternative policy which retrieves n documents without ensuring that their associated probabilities are in decreasing order can have its documents reordered to be made equivalent in retrieval effectiveness to one that does. Since the reordered policy is stochastically dominated by the standard policy, the standard policy stochastically dominates all retrieval policies. By our previous theorem, since the standard policy stochastically dominates all other policies, it also has optimal expected utility.

*Two sequences cannot be in different decreasing orders and still rank the entire collection. Thus, we assume that each ranking covers less than the entire collection.

To summarize, in this section we have seen that an inquirer may receive decreasing value from each relevant document he or she retrieves. Nevertheless, it is still advisable to follow the standard retrieval policy if the inquirer independently assesses the relevance of every document he or she retrieves and the IR system is well-calibrated and reports probabilities with certainty.

Comment: Stopping Rules, Utilities, and Costs

Our analysis has only considered "binomial" retrieval policies, which are policies whose stopping rule is that a total of n documents has been retrieved. However, other retrieval policies may dictate that a search conclude when a specified number of *relevant* documents has been retrieved. If two such "negative binomial" policies both retrieve r relevant documents, then they differ in effectiveness only in terms of the number of *non*relevant documents each has retrieved. Each of these nonrelevant documents imposes some cost on the inquirer.

The better of the two policies imposes the lower total cost. Intuitively, total costs should be lower when fewer nonrelevant documents are retrieved. Since the standard policy will ordinarily require the inquirer to examine fewer documents than any other policy, it imposes the least cost in most reasonable retrieval situations. Specifically, the standard policy imposes the lowest expected total cost whenever the cost associated with retrieving x nonrelevant documents is a linear or quadratic function of x.

We point out as well that our analysis ignores the exact order in which a sequence of relevant and nonrelevant documents has been retrieved. That is, two binomial retrieval sequences $S_1 = \langle \text{Rel, Nonrel, Nonrel, Rel, Rel} \rangle$ and $S_2 = \langle \text{Rel, Rel, Rel, Nonrel, Nonrel} \rangle$ are considered to have identical utility, since both sequences contain three relevant documents and the utility of each relevant document retrieved is a function of the number of relevant documents that has been retrieved already. Similarly, the analysis of cost functions for negative binomial retrieval policies is based solely on the number of nonrelevant documents and not their position within a temporal sequence of retrieved documents.

Additionally, the costs of retrieval (C_{11}, C_{12}, C_{21}, and C_{22}) depend only on whether a document is retrieved (or not) and is relevant (or not). We have not considered the different amounts of effort required to access documents located on different databases or the differences in connect-time costs or retrieval costs imposed by different systems.

Summary and Conclusion

To this point, we have considered retrieving documents by a strict rank-ordering of their probabilities of relevance and have seen this standard policy is defensible in the face of two possible objections. The first objection, a signal-detection characterization of IR behavior in conjunction with decision theoretic analysis, suggests that classification strictly by rank ordered retrieval status values is not always in order. The second, a utility theoretic analysis, indicates that expected utility does not necessarily increase with expected value, again suggesting that retrieval by strict rank ordering of probabilities of relevance may not be advisable.

In both instances, the standard policy has been rescued by mathematical analysis, but the analysis contains three important assumptions: One, the probability of relevance for any document is well-calibrated. Two, it is reported with certainty by the system. Three, an inquirer independently assesses the relevance of retrieved documents. We next examine these assumptions.

The assumption that an IR system produces well-calibrated probabilities, reported with certainty, means, in effect, that each document is assigned some predicted probability of relevance, p, by the system and, in the long run, p*100% of such documents will, indeed, be judged relevant by the user. Without this assumption, predictive probabilities of relevance behave like any other type of retrieval status value such as a simple matching score, Jaccard's similarity measure, etc. As we have seen, selecting for retrieval just those documents with greatest retrieval status values is not always theoretically defensible. Thus, without the assumptions of calibration and certainty, whether a document with a higher predictive probability of relevance should be retrieved before one with a lower one relies on empirical evidence. That is, we need to know the empirical *distribution* of predictive probabilities for both (1) relevant and (2) nonrelevant documents.

Similarly, the assumption that probabilities are well-calibrated and reported with certainty is necessary to demonstrate the optimality of the standard policy with respect to utilities. Without this assumption we cannot prove stochastic dominance, for we cannot even show that the probability of retrieving a relevant document with a single retrieval is greater for the standard policy than for an alternative policy, which we are required to show as the basis for our induction. Thus, with stochastic dominance in question, so is the optimality of the standard policy with respect to expected utility.

The assumption that the inquirer independently asseses the relevance of every retrieved document is necessary to show stochastic dominance. Without this assumption, the probability of relevance of an unretrieved document is a function of the set of previously retrieved documents and the relevance assessments made of them. This assumption rules out both the possibility of incorporating information describing how two documents are related with respect to their historical corelevance to the same queries and the possibility of learning from user feedback.

In summary, the standard retrieval policy, which agrees with the probabilistic ranking principle, has been generalized to include utilities and defended against possible anomalous behavior arising from signal-detection theory. This generalization and defense, however, rest on three assumptions. Alternative assumptions are possible which lead to modifications of the standard policy (Gordon & Lenk, in press).

References

Bookstein, A. (1977). When the most 'pertinent' document should not be retrieved—an anlaysis of the Swets model. *Information Processing and Management, 13,* 377–383.

Bookstein, A. (1985). Probability and fuzzy-set applications to information retrieval. In M. Williams, (ed.), *Annual Review of Information Science and Technology,* Vol 20., White Plains, NY: Knowledge Industry Publications.

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation, 19,* 30–41.

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.

DeGroot, M. & Feinberg, S. (1982). The comparison and evaluation of forecasters. *The Statistician, 32,* 12–22.

Gordon, M., & Lenk, P., in press. When is the probability ranking principle suboptimal? Working Paper, School of Business, University of Michigan.

Kahneman, D., Slovic, P., & Tverksy, A., (eds.) (1982). *Judgement under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

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

Swets, J. A. (1963). Effectiveness of information retrieval methods. *American Documentation, 20,* 72–89.