

When Information Retrieval Measures Agree About the Relative Quality of Document Rankings

Robert M. Losee

SILS, Manning Hall, CB#3360, University of North Carolina-Chapel Hill, Chapel Hill, NC 27599-3360.
E-mail: losee@ils.unc.edu

The variety of performance measures available for information retrieval systems, search engines, and network filtering agents can be confusing to both practitioners and scholars. Most discussions about these measures address their theoretical foundations and the characteristics of a measure that make it desirable for a particular application. In this work, we consider how measures of performance at a point in a search may be formally compared. Criteria are developed that allow one to determine the percent of time or conditions under which two different performance measures suggest that one document ordering is superior to another ordering, or when the two measures disagree about the relative value of document orderings. As an example, graphs provide illustrations of the relationships between precision and F .

1. Introduction

Measures of retrieval and filtering performance characterize different aspects of document orderings, usually showing better performance when relevant documents are moved toward the front of the ordered list of documents, and lower levels of performance when the relevant documents move away from the front of the list. Document ordering is usually performed by a retrieval or search engine using a ranking algorithm based on vector, probabilistic, or logical considerations. In some cases, two measurement techniques will differ on which of two orderings are superior. This work addresses the nature of these measurements and tries to show when the measurement induced orderings differ. We attempt to determine and understand those cases where different performance measures disagree about which document orderings are better than other orderings.

Information retrieval systems order documents based upon system-specific ranking criteria that take queries as indicators of the characteristics of documents to be retrieved

or ranked (Conlon & Conlon, 1998; Losee, 1998; Salton & McGill, 1983). Retrieval engines usually order documents based on criteria other than one of the popular measurement criteria. For example, few retrieval systems explicitly attempt primarily to maximize precision. The ordering of documents is often not monotonic with the values for many performance measures. Monotonicity may be understood informally as, given two sets of values, the condition that exists when an increase occurring in one set is accompanied by an increase in the other, although not necessarily by the same magnitude. Similarly, the values for documents ordered by one retrieval performance measure are often not monotonic with those of another performance measure.

We are interested in both the systems that order documents and the quality of the rankings produced by these systems. Information retrieval systems accept queries in a language consistent with the software of the system, search through indexes to find which documents best meet the information needs expressed in the query, and then retrieve these best documents from a set of stored documents. Internet search engines similarly accept queries and search through indexes, but the documents are not stored directly on the search engine's computers. The search engine provides an address specifying where on the Internet the document is located. Network filtering agents accept initial statements of information needs and then filter incoming documents for the user. Many of these types of systems functionally make the decision to retrieve or not retrieve, or ranks documents, based upon something like the probability the document is relevant.

Measures of retrieval performance are based on document orderings, the presentation of documents or media fragments in a particular order to the user. The documents are ordered by a particular variable or variables, and a study of the rational ordering of the documents may be based upon the distribution of the ordering variables over the set of all documents or the set of relevant documents (Losee, 1998). Distributions discussed below are such distributions.

We may study performance measures by formally describing those documents that have been retrieved and those

Received July 1, 1999; revised January 13, 2000; accepted January 13, 2000

© 2000 John Wiley & Sons, Inc.

that haven't been retrieved up to a specific point in the search. We can measure the proportion of documents that would be retrieved up to point x , assuming that we would want to work from the documents with the highest frequencies of occurrence for a desirable feature or feature complex down to documents with the occurrences with the lowest frequencies. Here x may be a single value for a characteristic, such as a term's frequency, or it may be interpreted as the value associated with the vector representing the set of characteristics being considered by the system.

Definition 1 [Survival Function $S(x, D)$]. The survival function $S(x, D)$ equals $1 - CDF(x, D)$, where $CDF(x, D)$ is the standard cumulative distribution function for the distribution D (Evans et al., 1993). For our purposes, distribution D is taken over the rank-ordered set of documents, with the survival function summing the probabilities of the characteristics in the range from x to the highest values that characteristics can hold (Losee, 1998).

The variable D denotes the distribution of the set of documents, ordered based on certain characteristics. The distribution of the ordered set of documents is such that density and cumulative distribution functions may be computed over the ordered list of documents.

2. Retrieval and Filtering Measures

Information retrieval performance is usually measured by considering to what degree documents relevant to the searcher are moved toward the front of the ordered list of documents. We refer the reader to other works that discuss more fully the issues associated with relevance (Iivonen & Sonnenwald, 1998; Ingwersen, 1996; Schamber, 1994; Sperber & Wilson, 1995; Tang & Solomon, 1998). Most measures depend on documents being assigned a binary relevance judgment. Continuous relevance scales can also be integrated into retrieval performance models (Losee, 1998). For purposes here, we assume that an individual can separate documents into two classes, which can be labeled as "relevant" and "nonrelevant." Common measures based upon relevance include measures commonly occurring in information retrieval research, such as precision and recall and precision-recall curves, as well as less common measures, such as normalized precision and recall, \mathcal{E} , F , expected search length and average search length. Additionally, performance measures may be based on concepts in addition to relevance, including document accuracy, novelty, currency, benefit, and system speed (Conlon & Conlon, 1998; Harter & Hert, 1997; Lancaster, 1968; Robertson & Hancock-Beaulieu, 1992; Schamber et al., 1990; Tague-Sutcliffe, 1995; Tang & Solomon, 1998).

The most commonly used performance measures are based upon *precision* and *recall*. The *recall* for a set of retrieved documents is the percent of relevant documents in the database that have been retrieved. The *precision*, given a set of retrieved documents, is the percent of documents that have been retrieved that are relevant. Searches may be categorized into two forms based upon precision and recall:

high-recall searches, in which most or all of the documents on a topic should be retrieved, and *high-precision* searches, in which the set of documents to be retrieved, often consisting of a small number of documents, is expected to be composed of predominately relevant documents.

Related to precision and recall are the E and *optimal- E* measures (Bollmann & Cherniavsky, 1981; Shaw, 1986; Van Rijsbergen, 1974). The E measure is computed as $E = 1 - 2/(P^{-1} + R^{-1})$, where P is the precision and R is the recall. The F measure is computed as $1 - E$ and has the attractive feature that high values represent better performance than lower values, the opposite of the situation with the E measure (Shaw et al., 1997).

Many retrieval measures determine performance at a point in the search process, such as the measures above, whereas other measures may compute performance based on the totality of the search process. The average search length (ASL) is the expected position of a relevant document in the ordered list of all documents (Losee, 1998). This is related to the expected search length (ESL), the number of nonrelevant documents retrieved up to a certain point in the search (Cooper, 1968), and the ranked half-life indicator, which represents the median position for the relevance frequency or relevance values for documents preceding a cutoff point (Borlund & Ingwersen, 1998). The ASL is measured in units of "documents"; knowing that the average position of a relevant document is 23 or 500 or 2 million documents into the ranked list of documents conveys useful information to the searcher, and may be much easier to interpret than a precision of .4 at a recall of .60 or an F of .837.

Several other measures have been used in retrieval studies, and the reader is referred to other sources for further discussions of individual performance measures (Bollmann & Cherniavsky, 1981; Cooper, 1973; Harter & Hert, 1997; Kanger, 1972; Losee, 1998; Robertson, 1969; Rorvig, 1988; Salton & McGill, 1983; Su, 1991; Tague & Farradane, 1978).

The performance measures analyzed as shown in the following section are those measures that are taken at a particular point in the search process. Thus, we address measures such as precision, E , F , and ESL. We exclude measures that address averages over a range of performance positions, such as average search length.

3. Characterizing the Differences Between Retrieval Measures

Understanding the relationships between two different measuring techniques, denoted as M_1 and M_2 , can be useful when studying performance results made with one or the other or both measures. Here, each measure M_i represents a function that acts as a single number measure of retrieval performance. The inputs to each measurement function are either (1) a document ordering, or related probability distribution over the set of documents, or (2) a single value, representing a parameter of the document ordering that is

used by the measure. Individual measures can be characterized theoretically (Pitts, 1972; Sutherland, 1975). We concern ourselves here with the relationships between such measures in retrieval and filtering contexts.

A number of concepts may prove useful when describing the relationships between measures. We begin by considering what it means to say that measures are equal.

Definition 2 (Equivalent Measures). When measures $M_i(D_v) = M_j(D_v)$ for all distributions of document orderings D_v and for the two measures M_i and M_j , then M_i and M_j are said to be equivalent measures. Similarly, when measure $M_i(D_v) \neq M_j(D_v)$ for some D_v , we say that measure M_i is different from measure M_j .

Considering the equivalence of measures is useful if we wish to decide whether two different formulae are the same. However, we may choose to focus more on the ordering of documents, which serves as the foundation for many notions of retrieval performance. This moves beyond simple numeric equivalence of measures to ordering equivalence.

Definition 3 (Equivalent Ordering Measures). Measures M_i and M_j are equivalent ordering measures, denoted as $M_i(D_v) \equiv M_j(D_v)$, when for all distributions of document sets D_v and D_w , $M_i(D_v) \geq M_i(D_w)$ if and only if $M_j(D_v) \geq M_j(D_w)$, and $M_i(D_v) < M_i(D_w)$ if and only if $M_j(D_v) < M_j(D_w)$.

Sometimes we will refer to measures over a given domain j of distributions, and determine whether two measures are equivalent ordering measures over domain j .

Measures such as the simple match and Jacquard measures are neither equivalent measures nor equivalent ordering measures. However, the Jacquard and Dice measures have been shown to be monotonic (Gower, 1971; Gregson, 1975; Sneath & Sokal, 1973). Thus, although they are not equivalent measures, they are equivalent ordering measures.

3.1. An Example of Nonequivalent Ordering Measures

Consider the case where we have two measures, $P_{.25}(x)$ and $P_{.75}(x)$, the precisions at the 25% and 75% recall levels. Each precision is taken immediately after the given recall level is obtained. Consider two different orderings of four relevant and six nonrelevant documents, with the documents being strongly ordered from left to right:

$X: NNRRRRNNNN$

and

$Y: NRNNRNNRRR$

Computing $P_{.25}(X)$ yields $\frac{1}{3}$ whereas $P_{.75}(X)$ yields $\frac{3}{5}$. Similarly, computing $P_{.25}(Y)$ yields $\frac{1}{2}$ whereas $P_{.75}(Y)$ yields $\frac{3}{9}$. In this case, the values for measure $P_{.25}(\)$ are ordered $X(\frac{1}{3})$ and then $Y(\frac{1}{2})$, whereas the values for performance measure $P_{.75}(\)$ are ordered $Y(\frac{3}{9})$ and then $X(\frac{3}{5})$. We can see

that $P_{.25}(\)$ and $P_{.75}(\)$ are not equivalent ordering measures over the set of document orders X and Y .

3.2. An Example of Equivalent Ordering Measures

A different set of document orderings is

$X': NRRRRNNNN$

and

$Y': NRNNRNNRRR$

Computing $P_{.25}(X')$ yields $\frac{1}{2}$ whereas $P_{.75}(X')$ yields $\frac{3}{5}$. Similarly, computing $P_{.25}(Y')$ yields $\frac{1}{2}$ whereas $P_{.75}(Y')$ yields $\frac{3}{5}$. The two measures, $P_{.25}(\)$ and $P_{.75}(\)$, equivalently order the document orderings X' and Y' . In addition, these two are equivalent measures over the set of orderings X' and Y' .

4. Determining Equivalent Ordering

It is useful to have a function that can indicate whether two measures are equal for a given document ordering, based on a set of parameters that govern the document ordering process. This leads us to the equivalent ordering function:

Theorem 1 (Equivalent Ordering Function). Given an equation similar in form to the definition for the equivalent ordering measures above, we find that $M_i \equiv M_j$ for all distributions of document orderings D_v and D_w when

$$f(M_i, M_j, D_v, D_w) = [Z(M_i(D_v) - M_i(D_w)) \\ = Z(M_j(D_v) - M_j(D_w))]$$

We denote the nonnegative function as $Z(x)$. The value for this function is either *true* or *false*. This theorem assumes that we can compute the nonnegative function, $Z(x)$, so that those values that are greater than or equal to zero have the value *true* and those that are negative have the value *false*. By computing the difference of the measured value at two points in the search for each of two measures and then comparing the nonnegativity of the difference associated with each measure, we compute whether the order of the two measures is the same for parameter sets or document orderings D_v and D_w . For our purposes, the magnitudes of the differences are unimportant.

Definition 4 (Measure Difference Function). We refer to

$$g(M_i, D_v, D_w) = Z(M_i(D_v) - M_i(D_w))$$

as the measure difference function for measure M_i and distributions of document orderings D_v and D_w .

This function returns either *true* or *false* depending on whether $M_i(D_v) \geq M_i(D_w)$ or not. This can be used to

examine the value of one side or the other of the equality in the equivalent ordering function. Given this, we may wish to consider the parameters that produce positive or negative values for this function.

5. Precision and Recall

We may compute the recall at point x in the ordering of documents as

$$R_x = S(x, \text{rel})$$

where x is the value for a feature or set of features. We denote rel and all as the document distributions D_{rel} and D_{all} for the sets of relevant, and of all, documents, respectively, with each document having a profile or feature vector, such as x , which is used to order (and retrieve) the documents. To allow for us to compare a wide range of measures, we express measures in terms of the survival function, $S(x, D_v)$, representing the proportion of events in the distribution of ordered documents D_v that occur at x or above (earlier in the retrieval process). We assume that documents are parameterized for our purposes so that retrieval consists of selecting documents in decreasing value of the probability density function of y [i.e., $\text{Pr}(y)$]. The function $S(x, \text{rel})$ has as its value the proportion of relevant documents with profiles at x or above.

To find the appropriate document profile x for a given level of recall, R , we can compute the inverse function

$$x = S^{-1}(R, \text{rel}) \quad (1)$$

An inverse function acts so as to provide the function which “undoes” the original operations, thus $x = f^{-1}(f(x))$ for all reversible functions. We find that this inverse survival function accepts as input the desired percent recall and produces the point on the distribution at which this occurs.

We may describe precision at document position (or profile) x_R as

$$P_{x_R} = \frac{S(x_R, \text{rel})}{S(x_R, \text{all})} \text{Pr}(\text{rel})$$

Using this, we may compute the precision for a given recall level R as

$$P_R = \frac{S(S^{-1}(R, \text{rel}), \text{rel})}{S(S^{-1}(R, \text{rel}), \text{all})} \text{Pr}(\text{rel})$$

Using our definition for precision and the measure difference function, we can determine the measure difference function for precision as

$$g(P_R, R_i, R_j) = Z(P_{R_i} - P_{R_j})$$

By expanding this and canceling out the $\text{Pr}(\text{rel})$ components, we arrive at

$$g(P_R, R_i, R_j) = Z \left(\frac{S(S^{-1}(R_i, \text{rel}), \text{rel})}{S(S^{-1}(R_i, \text{rel}), \text{all})} - \frac{S(S^{-1}(R_j, \text{rel}), \text{rel})}{S(S^{-1}(R_j, \text{rel}), \text{all})} \right) \quad (2)$$

6. Comparing Different Precision-Recall Measures

A single number measure of performance may be the single precision value, for example, at the 50% level of recall, denoted as $P_{.50}()$. We may numerically compute the value of this precision, given different input parameters. Additionally, we may determine the conditions under which two different precision measures yield the same values, either analytically or numerically. The differences owing to different situations may be understood as either different ordering distributions or, more parametrically, as the same family of distributions (e.g., the normal distribution, with different values for parameters, e.g., different means).

In information retrieval research, one often finds performance averaged over several different recall levels. Precision may be averaged over the 25%, 50%, and 75% recall levels to produce the $P_{.25,.50,.75}()$ value at point x :

$$P_{.25,.50,.75}(x) = \frac{P_{.25}(x) + P_{.50}(x) + P_{.75}(x)}{3}$$

Consider this in comparison to using the precision taken solely at 50% recall, $P_{.50}()$. Two differences are obvious. Moving from $P_{.25,.50,.75}()$ to $P_{.50}()$ removes both the $P_{.25}()$ and $P_{.75}()$ values and increases the weight placed on $P_{.50}()$. When comparing average precisions, there will always be the possibility of two kinds of changes: (1) the addition or deletion of specific P points, and (2) a change in the relative weightings for the P points.

We may compute the measure difference function as

$$g(P_{.25,.50,.75}(x,y)) = Z \left[\frac{S(P_{.25}(x), \text{rel})}{S(P_{.25}(x), \text{all})} - \frac{S(P_{.25}(y), \text{rel})}{S(P_{.25}(y), \text{all})} + \frac{S(P_{.50}(x), \text{rel})}{S(P_{.50}(x), \text{all})} - \frac{S(P_{.50}(y), \text{rel})}{S(P_{.50}(y), \text{all})} + \frac{S(P_{.75}(x), \text{rel})}{S(P_{.75}(x), \text{all})} - \frac{S(P_{.75}(y), \text{rel})}{S(P_{.75}(y), \text{all})} \right]$$

We can simplify $g(Z_{R,x,y})$ for the average precision value in the set of values set as

$$g(P_{\text{set}}(x,y)) = Z \left[\sum_{i \in \text{set}} \frac{S(P_i(x), \text{rel})}{S(P_i(x), \text{all})} - \frac{S(P_i(y), \text{rel})}{S(P_i(y), \text{all})} \right]$$

This general form of the equation may be used to analyze other averages of precision measures (Burgin, 1999). For example, the average precision taken at 11 recall levels, the precision at the 0%, 10%, 20%, . . . , 90%, 100% recall levels, captures the precision at a wide range of recall levels, and may be seen as better characterizing the performance distribution than a simpler measure, such as the precision taken at only the 50% point.

7. E and F Measures

The *E* and *F* measures provide single number measures of performance that have attractive theoretical bases and practical advantages (Bollmann & Cherniavsky, 1981; Shaw, 1986; Shaw et al., 1997; Van Rijsbergen, 1974). We may define *E* thus:

$$\begin{aligned} E &= 1 - \frac{2}{\frac{1}{P} + \frac{1}{R}} \\ &= 1 - \frac{2}{\frac{1}{\text{Pr}(\text{rel})S(x, \text{rel})} + \frac{1}{S(x, \text{rel})}} \\ &= 1 - \frac{2}{\frac{1 + S(x, \text{all})/\text{Pr}(\text{rel})}{S(x, \text{rel})}} \\ &= 1 - \frac{2S(x, \text{rel})}{1 + S(x, \text{all})/\text{Pr}(\text{rel})} \end{aligned}$$

Given this, and the fact that $F = 1 - E$, *F* may be defined thus:

$$F = \frac{2S(x, \text{rel})}{1 + S(x, \text{all})/\text{Pr}(\text{rel})}$$

Given this model, we can define the measure difference function $g(F)$ as

$$Z \left(\frac{2S(x, \text{rel})}{1 + S(x, \text{all})/\text{Pr}(\text{rel})} - \frac{2S(y, \text{rel})}{1 + S(y, \text{all})/\text{Pr}(\text{rel})} \right)$$

This is equivalent to

$$g(F) = Z \left[\frac{2 \text{Pr}(\text{rel})S(x, \text{rel})}{1 + S(x, \text{all})} - \frac{2 \text{Pr}(\text{rel})S(y, \text{rel})}{1 + S(y, \text{all})} \right] \quad (3)$$

8. Comparing Measure Difference Functions

We illustrate the use of these techniques by comparing the precision and *F* measures. Studying similarity using

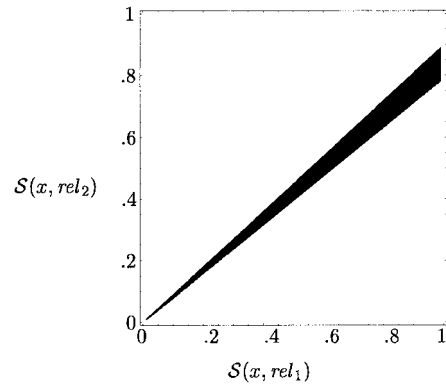


FIG. 1. Darkened region represents values for precision and *F* measures such that the equivalent measure function is false. Thus, the darkened areas represent where the measures have different preferences for the two document orderings, holding $S(x, \text{all}_1) = .95$, $S(x, \text{all}_2) = .75$, and $\text{Pr}(\text{rel}) = .02$ constant.

graphic tools provides a way to quickly and clearly visualize relationships; the same holds for studying relationships between retrieval measures. The graphs shown below were produced using Mathematica 4.0, a general mathematical processing and graphics package. For purposes here, we assume two distributions for the features used in ranking all documents and two distributions for the features used in ranking relevant documents. We arbitrarily assume that $S(x, \text{all}_1) = .95$, $S(x, \text{all}_2) = .75$, and that the generality, or unconditional probability that a document is relevant at, is .02.

Given these fixed values, we may find when the precision measure suggests the same ordering preference for documents as the *F* measure. In Figure 1, we can see that in most instances the precision measure prefers the same ordering as does the *F* measure, that is, where the equivalent ordering function is true. The exception to this is slightly below the diagonal, with differences occurring when $S(x, \text{rel}_1)$ is near $S(x, \text{rel}_2)$, with the difference growing as both values grow.

When the value for $S(x, \text{all}_2) = .40$, a graph such as that in Figure 2 is obtained.

A casual examination of these two figures suggests that under most circumstances these two measures will judge a pair of rankings as being in the same order (i.e., both will agree that one document is better or worse than the other document). The differences between measures occur when the orderings of documents are rather similar. This is consistent with saying that the two measures agree on gross differences but quibble over the fine points, which is what one would expect from measures both based on considerations of the same phenomena, such as precision and recall. Note that the figures should not be construed as implying that a certain percentage of time the two measures will be consistent in their preferences. The graphs show whether the measures agree on their preferences for a given set of values, as given on the *x* and *y* axes.

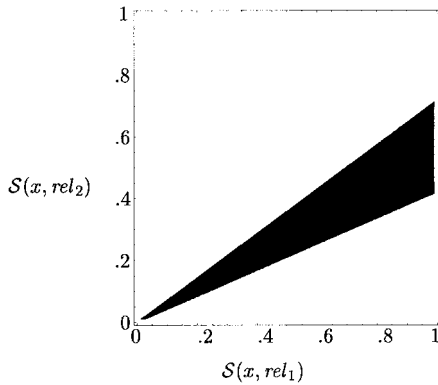


FIG. 2. Darkened region represents values for precision and F measures such that the equivalent measure function is false. Thus, the darkened areas represent where the measures have different preferences for the two document orderings, holding $S(x, all_1) = .95$, $S(x, all_2) = .40$, and $Pr(rel) = .02$ constant.

8.1. An Example of the Measure Difference Function for Precision and F Measures

A more detailed examination of the operation of the precision and F measures may produce a greater understanding of the data shown in Figures 1 and 2. Figures 3 and 4 show in darkened form those parameter values on the x and y axes for which the measure difference function is negative for precision and for F measures, respectively, holding other parameters as in Figure 2. The difference between these two figures (imagine overlaying Figures 3 and 4 and noting where they differ) can be seen as a broadening stripe across the lower to middle portion of the graph, is represented by the darkened area in Figure 2.

Equation (3) was used in producing Figure 4. There is a linear relationship between the value of the measure difference function for F and $S(x, rel_1)$. This suggests that the “clear” areas in the bottom right in Figures 3 and 4 may be defined as the set of points where

$$S(x, rel) \geq S(x, all) \frac{S(y, rel)}{S(y, all)}$$

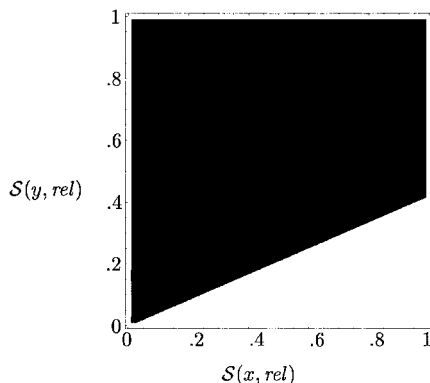


FIG. 3. Darkened region represents those values for precision such that the measure difference function is negative, using the parameters from Figure 2.

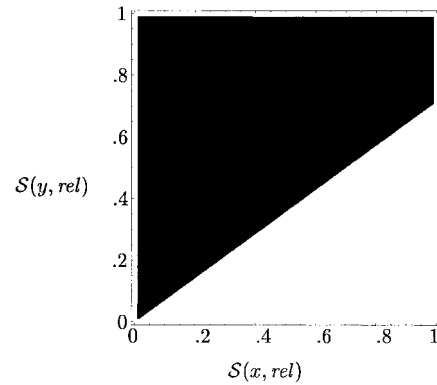


FIG. 4. Darkened region represents those values for F such that the measure difference function is negative, using the parameters from Figure 2.

When these survival functions are simple points, one can easily solve for the line where the measure difference function equals 0. Conversely, given a set of survival functions that are simple points, in many cases one can construct a measure function that is consistent with the points.

9. Discussion and Conclusions

Users and scholars presented with a retrieval system may want to apply methods such as these to determine in what circumstances or how often two different measures will have the same or different relative orderings, based on data from database retrievals. In environments with very high degrees of preference agreement between two measures, the choice of which of the two measures is chosen as the primary performance measure is less important. However, environments with lower degrees of measurement preference agreement place a greater burden on the user to understand the nature of the measures because a system that seems superior when evaluated using one measure may seem inferior using another measure. The user must interpret performance values produced by both measures appropriately.

There are many different measures of information retrieval and filtering performance. One measure may prefer different document orderings than a second measure because of document ordering characteristics and measure-specific concerns. We have developed criteria here for comparing these measures of retrieval performance up to a point in the retrieval process, and have proposed formal methods for determining when different measures would agree that one ranking is better than or the same as the other ranking. The measure difference function $g(\)$ for a particular measure has a value that, when compared to the $g(\)$ function for a different measure, returns *true* when the order preference is the same and *false* when the order preference is different. Using the measure difference function, we can study under what conditions increases in one measure’s value will correspond with an increase in the value of the other measure.

Being able to capture the relationships between measures is invaluable. They may be used to compare retrieval results

measured using different techniques. As an example, we show how the F measure and the precision (Z) measure provide the same or different relative values in particular environments. Being able to fully characterize and understand measures is at the heart of any field, and we feel that this work can help us to develop a greater understanding of retrieval and filtering theory and practice.

Acknowledgments

The author wishes to thank Dr. Robert Burgin of the North Carolina State Library for a discussion about Burgin (1999) that resulted in ideas leading to this article, as well as two anonymous referees for their suggestions.

References

- Bollmann, P., & Cherniavsky, V. (1981). Measurement-theoretical investigation of the MZ-metric. In R. Oddy, S.E. Robertson, C.J. van Rijsbergen, & P.W. Williams (Eds.), *Information retrieval research* (pp. 256–267). London: Butterworths.
- Borlund, P., & Ingwersen, P. (1998). Measures of relative performance and ranked half-life: Performance indicators for interactive IR. In W.B. Croft, A. Moffat, C.J. Van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia (pp. 324–331). New York: ACM Press.
- Burgin, R. (1999). The Monte Carlo method and the evaluation of retrieval system performance. *Journal of the American Society for Information Science*, 50, 181–191.
- Conlon, J.R., & Conlon, S.J. (1998). Robustness of well-designed retrieval performance measures under optimal user behavior. *Journal of the American Society for Information Science*, 49, 356–363.
- Cooper, W.S. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, 19, 30–41.
- Cooper, W.S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24, 87–100.
- Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions* (2nd ed.). New York: Wiley.
- Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857–874.
- Gregson, R.A.M. (1975). *Psychometrics of similarity*. New York: Academic Press.
- Harter, S.P., & Hert, C.A. (1997). Evaluation of information retrieval systems: Approaches, issues and methods. In Williams, M. (Ed.), *Annual review of information science and technology* (Vol. 32, pp. 1–94). Washington, D.C.: American Society for Information Science.
- Iivonen, M., & Sonnenwald, D.H. (1998). From translation to navigation of different discourses: A model of search term selection during the pre-online stage of the search process. *Journal of the American Society for Information Science*, 49, 312–326.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52, 3–50.
- Kanger, S. (1972). *Measurement: An essay in philosophy of science*. Theoria, 38, 1–44.
- Lancaster, F.W. (1968). *Information retrieval systems: Characteristics, testing, and evaluation*. New York: Wiley.
- Losee, R.M. (1998). *Text retrieval and filtering: Analytic models of performance*. Boston: Kluwer.
- Pitts, C.G.C. (1972). *Introduction to metric spaces*. Edinburgh: Oliver and Boyd.
- Robertson, S.E. (1969). The parametric description of retrieval tests: Part II: Overall measures. *Journal of Documentation*, 25, 93–107.
- Robertson, S.E., & Hancock-Beaulieu, M.M. (1992). On the evaluation of IR systems. *Information Processing and Management*, 28, 457–466.
- Rorvig, M.E. (1988). Psychometric measurement and information retrieval. In M.E. Williams (Ed.), *Annual review of information science and technology* (Vol. 23, pp. 157–189). Washington, D.C.: American Society for Information Science.
- Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Schamber, L. (1994). Relevance and information behavior. In Williams, M.E. (Ed.), *Annual review of information science and technology* (Vol. 29, pp. 3–48). Washington, D.C.: American Society for Information Science.
- Schamber, L., Eisenberg, M., & Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26, 755–776.
- Shaw, Jr., W.M. (1986). On the foundation of evaluation. *Journal of the American Society for Information Science*, 37, 346–348.
- Shaw, Jr., W.M., Burgin, R., & Howell, P. (1997). Performance standards and evaluations in IR test collections: Vector-space and other retrieval models. *Information Processing and Management*, 33, 15–36.
- Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy: The principles and practices of numerical classification*. San Francisco: W.H. Freeman.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell.
- Su, L.T. (1991). *An Investigation to Find Appropriate Measures for Evaluating Interactive Information Retrieval*. PhD thesis, Rutgers, New Brunswick, NJ.
- Sutherland, W.A. (1975). *Introduction to metric and topological spaces*. Oxford: Clarendon Press.
- Tague, J., & Farradane, J. (1978). Estimation and reliability of retrieval effectiveness measures. *Information Processing and Management*, 14, 1–16.
- Tague-Sutcliffe, J. (1995). *Measuring information: An information services perspective*. San Diego: Academic Press.
- Tang, R., & Solomon, P. (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing and Management*, 34, 237–256.
- Van Rijsbergen, C. (1974). Foundation of evaluation. *Journal of Documentation*, 30, 365–373.