Measuring Retrieval Effectiveness Based on User Preference of Documents

Y.Y. Yao

Department of Mathematical Sciences
Lakehead University
Thunder Bay, Ontario
Canada P7B 5E1

Abstract

The notion of user preference is adopted for the representation, interpretation and measurement of the relevance or usefulness of documents. User judgments on documents may be formally described by a weak order (i.e., user ranking) and measured using an ordinal scale. Within this framework, a new measure of system performance is suggested based on the distance between user ranking and system ranking. It only uses the relative order of documents and therefore confirms to the valid use of an ordinal scale measuring relevance. It is also applicable to multi-level relevance judgments and ranked system output. The appropriateness of the proposed measure is demonstrated through an axiomatic approach. The inherent relationships between the new measure and many existing measures provide further supporting evidence.

1 Introduction

Representation of user judgments on the usefulness of documents is fundamental to the development of reliable methods and techniques for measuring the effectiveness of information retrieval systems. Typically, user judgments are described using the concept of *relevance* (Bookstein, 1979; Cooper, 1971; Cuadra & Katter, 1967; Saracevic, 1970, 1975). A document is relevant if the user considers the document to be useful, otherwise it is non-relevant. Based on such dichotomous decisions, a number of performance measures have been proposed, such as precision, recall, fallout, normalized precision and recall, and the expected search length (Bollmann *et al.*, 1992; Cooper 1968; Robertson, 1969; Salton, 1992; Salton & McGill, 1983; Sparck Jones, 1981; van Rijsbergen, 1974, 1979). However, the use of the dichotomous notion of relevance has led to some criticism of these measures (Janes, 1991; Robertson, 1969; Saracevic, 1970). The insufficient treatment of the ranking effect of a retrieval system in the standard measures of precision and recall has also been questioned by many authors (Cooper, 1968; Raghavan, Bollmann, & Jung, 1989; Robertson, 1969). Although some research has been done on the evaluation of retrieval performance, using a multi-valued relevance scale (Bollmann et. al, 1986; Frei & Schäuble, 1991; Keen, 1971; Rocchio, 1971), there is still a lack of rigorous investigation on this important topic.

One difficulty in using a multi-valued relevance scale is that there is no clear guidance of how to design such a scale. Various proposals have been suggested, ranging from numeric quantification to verbal qualification. For example, Keen (1971) used a four-valued scale called *point grade* which assigns four to the most relevant documents, three to the next, and two and one to the final two groups. Saracevic *et al.* (1987) used a three-valued scale consisting of *relevant*, *partially relevant*, and *non-relevant*. Maron and Kuhns (1970) adopted an even finer relevance scale consisting of *very relevant*, *relevant*, *somewhat relevant*, *only slightly relevant* and *non-relevant*. Since a universal interpretation for these multi-valued relevance scales does not appear to exist, the compatibility between them is not entirely clear. If the numbers or the

verbal descriptions are not fully understood, a multi-valued relevance scale may be easily misused in relevance assessment and system evaluation (French, 1986; King, 1968). For instance, there may exist a potential problem in using Keen's revised precision and recall measures. On the one hand, the *addition* of relevance grades suggests that a document with grade two is equivalent to two documents with grade one. On the other hand, a user may not be aware of this implication when assessing the relevance grade for each document. The same observation is also true for the calculation of the average of rank used by Rocchio (1971). The investigation of Eisenberg and Barry (1988) further indicated that the adoption of a finite and fixed multi-valued relevance scale may be inappropriate for measuring and reflecting user judgments. They suggested that an open-ended scale (i.e., magnitude estimation) should be used in which the user is not constrainted to use a prefixed set of relevance values.

In this paper, the concept of user preference is adopted from decision and measurement theories as a primitive notion. Within this framework, instead of stating whether a document is relevant or not, a user specifies whether a document is more, or less relevant than another document, i.e., whether the user prefers one document to another (Wong, Yao, & Bollmann, 1988). One of our goals is to establish a basis for the study of the representation, interpretation and measurement of user judgments on the usefulness of documents, using a multi-valued relevance scale. A list of properties is explicitly stated regarding user preference on documents. Under this set of properties, a user in fact provides a ranked list of documents. Any measurement of relevance is therefore based on an ordinal scale. This interpretation of relevance has an important impact on the design of system performance measures. For example, the use of the absolute values as in Keen's revised precision and recall measures may not necessarily be meaningful, as they are not invariant to strictly monotonic increasing transformations of the absolute values.

Another objective of this paper is to propose a new system performance measure. Since many experimental retrieval systems provide a ranked list of documents, a

measure of effectiveness may be defined by examining the agreement or disagreement between the user and the system rankings. For this purpose, it is suggested that a normalized version of Kemeny and Snell's distance function between rankings is used. This new measure only uses the information about the relative order of documents in a ranking. It confirms to the valid use of an ordinal scale. The appropriateness of the proposed performance measure is demonstrated through an axiomatic approach. The rationale behind the axioms on the distance function is explained in the context of information retrieval. Supporting evidence is further provided by showing the relationships between the new measure and many existing measures.

2 Representation and Measurement of User Judgments on Documents

The notion of user preference has been discussed in the literature of information retrieval, although its usefulness has perhaps not been fully explored. It was implicitly used in the work of Cooper (1971) to differentiate two different interpretations of relevance: relevance as a logical relation between documents and queries, and relevance (or its judgment by a user) as a utility or significance judgment. Kochen (1974) formally defined the notion of utility in terms of user preference. Robertson and Sparck Jones (1976) pointed out that user preference may be used as a basis for computing term weights. Bookstein (1983, 1989) suggested using the preference structure to estimate the expected cost of retrieving a document or a set of documents in a set-oriented retrieval framework. Based on measurement theory, Bollmann and Wong (1987) discussed the necessary and sufficient conditions on the user preference for the justification of using linear functions in many retrieval models. Similarly, Wong, Bollmann and Yao (1991) attempted to establish a measurement-theoretic foundation for information retrieval. Based on the results of these studies, the concept of user preference is adopted for the measurement of relevance in the present investigation.

The user preference on documents can be described through a pairwise comparison of documents. Given two documents in a collection, it is assumed that a user is able

to decide if one document is more useful or relevant than the other document. Let D denote a finite set of documents. The user preference may be formally defined by a binary relation $\succ$ on D: for $d, d' \in D$,

$$d \succ d' \Longleftrightarrow \text{the user prefers } d \text{ to } d'. \tag{1}$$

The relation $\succ$ is called a (strict) preference relation, which is a subset of the Cartesian product $D \times D$:

$$\succ \ = \{(d, d') \mid \text{the user prefers } d \text{ to } d'\}. \tag{2}$$

Using a preference relation, a user only provides the relative relevance judgments on documents without referring to any predefined relevance scale. If $d \succ d'$ holds, d is said to be preferred to d'. This may be paraphrased as d being more useful or relevant than d' (Rocchio, 1971; Wong & Yao, 1990; Wong, Yao, & Bollmann, 1988). In the absence of strict preference, i.e., if both $\neg(d \succ d')$ and $\neg(d' \succ d)$ hold, d is said to be indifferent to d'. An indifference relation $\sim$ on D can be defined as follows:

$$d \sim d' \Longleftrightarrow (\neg(d \succ d'),\ \neg(d' \succ d)). \tag{3}$$

The indifference relationship between documents d and d' may have several interpretations (Wong & Yao, 1990). A user may consider d and d' to be equally useful. Documents d and d' may be incomparable because it does not make sense to compare them from the user point of view. This situation may occur when a user is asked to judge between two documents which are both entirely irrelevant to the user's information needs.

We assume that the user reviews the entire set before making any ranking decision. The user ranking may be considered as the optimal arrangement of documents that is most beneficial to the user. There are many factors that may affect the user preference. For example, user preference may be influenced by the order in which documents are presented (Eisenberg & Barry, 1988). For simplicity, no attempt is made to define preference in terms of other concepts in this paper. A preference relation may have either an objective or a subjective interpretation, as in the case of

relevance. The discussion of performance measures is independent of any particular interpretation.

User judgments with some of the existing relevance scales may be easily interpreted in terms of user preference. Suppose $D_1, D_2, \ldots, D_m$ are subsets of documents being arranged in decreasing order of relevance under a m-valued relevance scale (Robertson, 1969; Rocchio, 1971). Such user judgments may be equivalently represented by a preference relation defined as:

$$d \succ d' \Longleftrightarrow \exists i, j (i < j, d \in D_i, d' \in D_j). \tag{4}$$

The corresponding indifference relation is given by:

$$d \sim d' \Longleftrightarrow \exists i (d \in D_i, d' \in D_i). \tag{5}$$

When $m = 2$, the standard two-valued relevance scale is obtained. In this case, the set of document D is divided into two disjoint subsets, the set of relevant documents rel and the set of non-relevant documents $nrel$. An equivalent representation, using a preference relation, is given by:

$$d \succ d' \Longleftrightarrow (d \in rel, d' \in nrel). \tag{6}$$

With this definition, the indifference relation is:

$$d \sim d' \Longleftrightarrow (d, d' \in rel) \text{ or } (d, d' \in nrel). \tag{7}$$

For the three-valued relevance scale used by Saracevic *et al.* (1987), a preference relation may be defined such that the *relevant* documents are preferred to *partially relevant* documents, which in turn are preferred to *non-relevant* documents.

From a measurement-theoretic point of view, it is important to identify the desired properties that a preference relation must satisfy, and determine if the preference can be measured using a particular relevance scale (French, 1986; Roberts, 1979). Consider the following two axioms:

$$\begin{aligned} \text{Asymmetry}: &\quad d \succ d' \Longrightarrow \neg(d' \succ d), \\ \text{Negative transitivity}: &\quad (\neg(d \succ d'), \neg(d' \succ d'')) \Longrightarrow \neg(d \succ d''). \end{aligned}$$

The asymmetry axiom requires that a user cannot prefer d to d' and at the same time prefers d' to d. The negative transitivity axioms states that if a user does not prefer d to d', nor d' to d'', the user should not prefer d to d''. A preference relation satisfying these two axioms is called a *week order*. If a preference relation is a weak order, it is transitive, i.e., $d \succ d'$ and $d' \succ d''$ imply $d \succ d''$. In the context of information retrieval, it seems reasonable that a user preference relation should satisfy these two properties. In a subsequent discussion, the preference relations that are weak orders will be considered.

A few additional properties of a weak order are summarized in the following lemma (Fishburn, 1970).

Lemma 1 Suppose a preference relation $\succ$ on a finite set of documents D is a weak order. Then,

a. the relation $\sim$ is an equivalence relation,

b. exactly one of $d \succ d', d' \succ d$ and $d \sim d'$ holds for every $d, d' \in D$,

c. the relation $\succ'$ on $D/\sim$ defined by

$X \succ' Y \Longleftrightarrow \exists d, d'(d \succ d', d \in X, d' \in Y)$,

is a linear order, where X and Y are elements of $D/\sim$.

A linear order is a weak order in which any two different elements are comparable. This lemma implies that if $\succ$ is a weak order, the indifference relation $\sim$ divides the set of documents into disjoint subsets. Furthermore, for any two equivalence classes X and Y of $\sim$, either $X \succ' Y$ or $Y \succ' X$ holds. In other words, if a preference relation $\succ$ is a weak order, it is possible to arrange the documents into several levels so that documents in a higher level are preferred to documents in a lower level, and documents in the same level are indifferent (Cooper, 1968). For convenience, a user preference relation is also called a user ranking.

In the measurement-theoretic terminology, the requirement of a weak order indeed suggests the use of a multi-valued relevance scale, as shown by the following representation theorem (Fishburn, 1970; Roberts, 1979).

Theorem 1 Suppose D is a finite non-empty set of documents and $\succ$ a relation on D. There exists a real-valued function $u : D \longrightarrow \mathbf{R}$ satisfying the condition,

$$d \succ d' \Longleftrightarrow u(d) > u(d') \tag{8}$$

if and only if $\succ$ is a weak order. Moreover, u is uniquely defined up to a strictly monotonic increasing transformation.

The numbers $u(d), u(d'), \ldots$ as ordered by $>$ faithfully reflect the order of $d, d', \ldots$ under $\succ$. The function u is referred to as an order-preserving utility function. It quantifies a user preference relation and provides a measurement of user judgments. The quantity $u(d)$ may be interpreted as the relevance value of document d. According to Theorem 1, the axioms of a weak order are the conditions which allow the measurement. There are two ways to view these axioms (Fishburn, 1970). The *prescriptive* or *normative* interpretation is concerned with the principles that a user must follow to specify a preference relation. The axioms are looked upon as conditions of rationality. A *rational* user's judgments must allow the measurement in terms of a quantitative utility function. On the other hand, the *descriptive* interpretation treats the axioms as testable conditions. Whether can measure the user judgments depends on whether the user preference relation is a weak order.

In the following corollary, Theorem 1 is extended to situations where the user judgments are measured, using a predefined multi-valued relevance scale.

Corollary 1 Let $\mathbf{L}$ be a linearly ordered set with a binary relation $>$. There exists a mapping $u : D \longrightarrow \mathbf{L}$ satisfying the condition,

$$d \succ d' \Longleftrightarrow u(d) > u(d') \tag{9}$$

if and only if $\succ$ is a weak order and $|D/\!\sim| \leq |\mathbf{L}|$, i.e., the number of equivalent classes in $D/\!\sim$ is less than or equal to the number of elements in $\mathbf{L}$.

When using a multi-valued relevance scale, it is necessary to have different values from $\mathbf{L}$ for distinct equivalence classes of $D/\!\sim$. The additional condition $|D/\!\sim| \leq |\mathbf{L}|$

is therefore needed in the corollary. According to the corollary, the three-valued relevance scale { *relevant, partially relevant, non-relevant* } cannot be used to measure a preference relation that induces four equivalence classes.

Example 1 Suppose a user preference relation $\succ$ on $D = \{d_1, d_2, d_3, d_4\}$ is specified by the following weak order:

$$d_3 \succ d_1, \quad d_4 \succ d_1, \quad d_3 \succ d_2, \quad d_4 \succ d_2, \quad d_4 \succ d_3.$$

The corresponding indifference relation $\sim$ has three equivalence classes $\{d_4\}$, $\{d_3\}$, and $\{d_1, d_2\}$. According to $\succ'$, these equivalence classes may be arranged into three levels:

$$\{d_4\} \succ' \{d_3\} \succ' \{d_1, d_2\}.$$

Obviously, the utility function defined by:

$$u_1(d_1) = 0, \quad u_1(d_2) = 0, \quad u_1(d_3) = 1, \quad u_1(d_4) = 2,$$

provides a measurement of $\succ$. To serve the same purpose, another utility function may be used:

$$u_2(d_1) = -1, \quad u_2(d_2) = -1, \quad u_2(d_3) = 1, \quad u_2(d_4) = 4.$$

Let $\mathbf{L} = \{$ *relevant, partially relevant, non-relevant* $\}$. The preference relation $\succ$ may also be measured, using a non-numeric utility function:

$$\begin{array}{ll} u_3(d_1) = \textit{non–relevant}, & u_3(d_2) = \textit{non–relevant}, \\ u_3(d_3) = \textit{partially relevant}, & u_3(d_4) = \textit{relevant}. \end{array}$$

Although, u_1, u_2 and u_3 use different absolute values, all of them preserve the same relative order for any pair documents. □

In the above discussion, only the asymmetry and negative transitivity axioms are considered, and consequently the ordinal scale for the measurement of user judgments. A crucial question is whether an ordinal scale is appropriate. In reviewing the

experimental results of Cuadra and Katter (1967), and Rees and Schultz (1967) on relevance judgments, Saracevic (1970) stated an important conclusion, namely that although the ratings of the degree of relevance by different users may be scattered, the *relative* position of documents as to their relevance may be expected to be remarkably consistent. The same results were also reported in a study by Lesk and Salton (1968). Such evidence indeed suggests that an ordinal scale should be used. There are also many advantages in using an ordinal scale. The notion of weak order is rich enough to represent any use judgments that can be expressed using a multi-valued relevance scale. The axioms of a weak order may be easily explained so that a user is guided in making preference assessment. There are no such restrictions as that only a limited number of levels can be used in expressing the user judgments with a predefined multi-valued relevance scale. This scheme is compatible with the open-ended scale for measuring relevance adopted by Eisenberg (1988), and Eisenberg and Barry (1988).

One important implication of Theorem 1 is that if only the two axioms of weak order are required, the user preference is measured by an *ordinal* utility function. For an ordinal scale, it is meaningful to examine the order induced by the utility function. In other words, comparison is a valid operation. Other arithmetic operations, such as addition and subtraction, are not necessarily meaningful (French, 1986). This observation is very crucial for the design of a system performance measure. One must be cautious when using an arithmetic operation on the absolute values of the utility function that measure the relevance of documents. Under the constraint of measuring relevance with an ordinal scale, we propose and examine a new system performance measure based on the distance between system and user rankings.

3 Distance Between Rankings

This section reviews the axiomatic approach used by Kemeny and Snell (1962) for the definition of a distance function between two rankings that are weak orders. By examining rationality of the required list of axioms in the context of information retrieval, it is argued that the distance function is suitable for the evaluation of a

retrieval system.

Before stating the axioms for distance between two rankings, a few important concepts are first introduced. Given a ranking $\succ$ on a set of documents D, the *restriction* of $\succ$ on a subset $S \subseteq D$ is defined by:

$$\succ(S) = \ \succ \cap (S \times S) = \{(d, d') \mid d \succ d' \text{ and } d, d' \in S\}. \tag{10}$$

Two rankings *agree* on a pair of documents $d, d' \in D$ if both of them rank d and d' in the same order, i.e., $\succ_1(\{d, d'\}) \ = \ \succ_2(\{d, d'\})$. They *contradict* on $\{d, d'\}$ if one ranking puts d higher and the other ranking puts d' higher. They are *compatible* on $\{d, d'\}$ if one ranking puts d or d' higher and the other ranking has d' and d tied. A ranking $\succ_2$ is *between* two rankings $\succ_1$ and $\succ_3$, written $B(\succ_1, \succ_2, \succ_3)$, if for each pair of documents d and d', $\succ_2(\{d, d'\})$ is between $\succ_1(\{d, d'\})$ and $\succ_3(\{d, d'\})$. That is, if $\succ_1$ and $\succ_3$ agree on $\{d, d'\}$, then $\succ_2$ must agree with them. If $\succ_1$ and $\succ_3$ are compatible on $\{d, d'\}$, then $\succ_2$ must agree with either $\succ_1$ or $\succ_3$. In the case that $\succ_1$ and $\succ_3$ contradict on $\{d, d'\}$, $\succ_2$ must declare a tie of d and d'.

A subset S of D is a *segment* in a ranking $\succ$ if every element $d \in D - S$ is either above every element of S or below every element of S. If $S \neq D$, S is a *proper segment*. Let $\overline{S}_\succ$ denote the set of documents above S under $\succ$, and $\underline{S}_\succ$ the set of documents below S under $\succ$. They are also segments. The rankings $\succ(S)$, $\succ(\overline{S}_\succ)$, and $\succ(\underline{S}_\succ)$ are the restrictions of $\succ$ on the segments S, $\overline{S}_\succ$, and $\underline{S}_\succ$. Consider two rankings $\succ_1$ and $\succ_2$ and a common segment S. It may be said that $\succ_1$ and $\succ_2$ agree *outside* S if $\overline{S}_{\succ_1} = \overline{S}_{\succ_2}$, $\underline{S}_{\succ_1} = \underline{S}_{\succ_2}$, $\succ_1(\overline{S}_{\succ_1}) \ = \ \succ_2(\overline{S}_{\succ_2})$, and $\succ_1(\underline{S}_{\succ_1}) \ = \ \succ_2(\underline{S}_{\succ_2})$.

Let $\Gamma(D)$ denote the set of all rankings over the set of documents D. A distance measure between rankings is a real-valued function $\beta : \Gamma(D) \times \Gamma(D) \longrightarrow \mathbf{R}$. It seems reasonable that a distance function should satisfy the following axioms (Kemeny & Snell, 1962; Roberts, 1976): for all $\succ_1, \succ_2, \succ_3 \in \Gamma(D)$,

Axiom 1.1 $\beta(\succ_1, \succ_2) \geq 0$, with quality if and only if $\succ_1 \ = \ \succ_2$.

Axiom 1.2 $\beta(\succ_1, \succ_2) = \beta(\succ_2, \succ_1)$.

Axiom 1.3 $\beta(\succ_1, \succ_2) + \beta(\succ_2, \succ_3) \geq \beta(\succ_1, \succ_3)$, with equality if and only if

$B(\succ_1, \succ_2, \succ_3)$.

Axiom 2 If ranking $\succ_1'$ results from ranking $\succ_1$ by a permutation of the set D and $\succ_2'$ results from $\succ_2$ by the same permutation, then $\beta(\succ_1, \succ_2) = \beta(\succ_1', \succ_2')$.

Axiom 3 Suppose $\succ_1, \succ_2, \succ_1'$ and $\succ_2'$ are rankings with a common segment S. If

$$\begin{aligned} &\text{(i)} \quad \succ_1 \text{ and } \succ_2 \text{ agree outside } S, \\ &\text{(ii)} \quad \succ_1' \text{ and } \succ_2' \text{ agree outside } S, \text{ and} \\ &\text{(iii)} \quad \succ_1(S) = \succ_1'(S) \text{ and } \succ_2(S) = \succ_2'(S), \end{aligned}$$

then $\beta(\succ_1, \succ_2) = \beta(\succ_1', \succ_2')$.

Axiom 4 The minimum postive distance between elements in $\Gamma(D)$ is 1, i.e., for all $\succ_1, \succ_2 \in \Gamma(D)$, $\beta(\succ_1, \succ_2) = 0$ or $\beta(\succ_1, \succ_2) \geq 1$, and for some $\succ_1, \succ_2 \in \Gamma(D)$, $\beta(\succ_1, \succ_2) = 1$.

Axioms 1.1-1.3 are the usual properties of any distance function. Axiom 2 states that a measure of distance does not depend on the particular objects chosen for ranking. That is, a distance measure is independent of how we label the documents in the collection. A relabeling of the set of documents does not affect the distance. For example, the distance between $d_1 \succ_1 d_2 \succ_1 d_3$ and $d_3 \succ_2 d_2 \succ_2 d_1$ is the same as the distance between $d_2 \succ_1' d_3 \succ_1' d_1$ and $d_1 \succ_2' d_3 \succ_2' d_2$, since $\succ_1'$ and $\succ_2'$ can be obtained from $\succ_1$ and $\succ_2$ by changing d_1 to d_2, d_2 to d_3, and d_3 to d_1. Axiom 3 suggests that if two rankings are the same at both the top and bottom, and differs only on a set of documents in the middle, then the distance between these two rankings should only depend on the ranking of the documents in the middle. Axiom 4 is introduced only for the sake of choosing a unit of measurement.

Axioms 1-4 are consistent. They are both necessary and sufficient for the existence of a unique distance measure (Kemeny & Snell, 1962; Roberts, 1976).

Theorem 2 For every finite set D with two or more members, there is a distance function β on $\Gamma(D) \times \Gamma(D)$ which satisfies axioms 1-4. Moreover, β is uniquely determined.

Suppose $\succ_1$ and $\succ_2$ are two rankings on D. A distance function satisfying axioms 1-4 can be constructed as follows. First, define the distance between rankings with respect to a pair of documents $d, d' \in D$. Let $\delta_{\succ_1,\succ_2}(d, d')$ count 0 if $\succ_1$ and $\succ_2$ agree on d and d', let $\delta_{\succ_1,\succ_2}(d, d')$ count 1 if $\succ_1$ and $\succ_2$ are compatible on d and d', and let $\delta_{\succ_1,\succ_2}(d, d')$ count 2 if $\succ_1$ and $\succ_2$ contradict on d and d'. The overall distance between two rankings can be calculated by:

$$\beta(\succ_1, \succ_2) = \sum_{d,d'} \delta_{\succ_1,\succ_2}(d, d'), \tag{11}$$

where the summation is over all unordered document pairs. Clearly, the computation of the distance between two rankings only depends on the relative order of documents.

Example 2 Consider the following two rankings on the set of documents $\{d_1, d_2, d_3, d_4\}$:

$$d_1 \succ_1 d_2 \succ_1 \begin{matrix} d_3 \\ d_4 \end{matrix} ,$$

$$d_2 \succ_2 \begin{matrix} d_1 \\ d_3 \end{matrix} \succ_2 d_4 .$$

For these two rankings, they yield:

$$\delta_{\succ_1,\succ_2}(d_1, d_2) = 2, \quad \delta_{\succ_1,\succ_2}(d_1, d_3) = 1, \quad \delta_{\succ_1,\succ_2}(d_1, d_4) = 0,$$
$$\delta_{\succ_1,\succ_2}(d_2, d_3) = 0, \quad \delta_{\succ_1,\succ_2}(d_2, d_4) = 0, \quad \delta_{\succ_1,\succ_2}(d_3, d_4) = 1.$$

From these values, the distance between $\succ_1$ and $\succ_2$ is given by:

$$\beta(\succ_1, \succ_2) = 2 + 1 + 0 + 0 + 0 + 1 = 4. \quad \square$$

4 Distance-based Measures of Retrieval Effectiveness

Normally, the result of a retrieval process is a ranked list of documents which is a weak order (Cooper, 1968). Let $\succ_u$ denote the user ranking and $\succ_s$ the system ranking. In the ideal situation, a retrieval system is expected to produce the user

ranking, i.e., $\succ_s = \succ_u$. This requirement is known as the *perfect ranking* criterion. A much weaker criterion may also be used in which a system is only required to rank preferred documents higher than the nonpreferred ones (Wong & Yao, 1990, Wong, Yao, & Bollmann, 1988). Such an acceptable ranking can be derived by arbitrarily rearranging the documents in the same equivalence class. Under this *acceptable ranking* criterion, the system performance is evaluated independent of how the system ranks the documents in the same equivalence class. Depending on the criterion adopted, the system performance may be measured in terms of the *divergence* of $\succ_s$ from, or the closeness of $\succ_s$ to, the ideal ranking $\succ_u$ or an acceptable ranking of $\succ_u$. In other words, we assume that a system producing a ranking closer to the user ranking is better than another system producing a ranking further away.

For a given ranking $\succ$, the *converse* ranking $\succ^c$ is defined by:

$$\succ^c = \{(d', d) \mid d \succ d'\}. \tag{12}$$

That is, the converse ranking can be obtained by reading the original ranking backward. Based on the notion of converse ranking and the perfect ranking criterion, the numbers of agreeing pairs C^+, contradictory pairs C^-, and compatible pairs C^0 for two rankings $\succ_u$ and $\succ_s$ are given by:

$$\begin{aligned} C^+ &= \mid \succ_u \cap \succ_s \mid, \\ C^- &= \mid \succ_u \cap \succ_s^c \mid = \mid \succ_u^c \cap \succ_s \mid, \\ C^0 &= \mid \succ_u \cap \sim_s \mid + \mid \sim_u \cap \succ_s \mid = C^u + C^s, \end{aligned} \tag{13}$$

where $\mid \cdot \mid$ denotes the cardinality of a set. From equation (11), the distance between $\succ_u$ and $\succ_s$ can be computed by the formula:

$$\beta(\succ_u, \succ_s) = 2C^- + C^0 = 2C^- + C^u + C^s. \tag{14}$$

Wong, Yao and Bollmann (1988) argued that the acceptable ranking criterion may be more suitable for information retrieval. Rocchio (1971) explicitly stated that the objective of a retrieval system is to produce an acceptable ranking. Many performance

measures such as precision, recall, normalized precision, and expected search length are in fact based on this criterion. This would suggest that a performance measure may be derived by using the distance between $\succ_s$ and an acceptable ranking of $\succ_u$. There are many acceptable rankings with respect to $\succ_u$. From the point view of effectiveness, all these acceptable rankings are equivalent. For the definition of a fair measure, one should choose an acceptable ranking closest to $\succ_s$. Let $\Gamma_u(D)$ denote the set of all acceptable rankings of $\succ_u$. The following distance-based performance measure (dpm) is suggested:

$$dpm(\succ_u, \succ_s) = \min_{\succ \in \Gamma_u(D)} \beta(\succ, \succ_s). \tag{15}$$

Based on the properties of the distance function, if $\succ_s$ is an acceptable ranking, then $dpm(\succ_u, \succ_s) = 0$.

Let $\succ_a$ denote an acceptable ranking closest to $\succ_u$. The definition of the distance function implies that $\delta_{\succ_a, \succ_s}(d, d')$ is minimum for every pair of documents. According to its definition, an acceptable ranking of $\succ_u$ may be obtained by rearranging the documents in the same equivalent class. For two documents with $d \sim_u d'$, we must have $\succ_a(\{d, d'\}) = \succ_s(\{d, d'\})$, for otherwise $\delta_{\succ_a, \succ_s}(d, d')$ is not minimum. Therefore, the $\succ_a$ can be constructed by:

$$\succ_a \;=\; \succ_u \cup (\sim_u \cap \succ_s). \tag{16}$$

Moreover, $\succ_a$ is the only acceptable ranking closest to $\succ_s$. Consequently, the distance-based performance measure can be equivalently defined by:

$$dpm(\succ_u, \succ_s) = \beta(\succ_a, \succ_s). \tag{17}$$

The computation of pbm can be easily carried out as follows. For rankings $\succ_a$ and $\succ_s$, the number of agreeing, contradictory, and compatible pairs are:

$$\begin{aligned} |\succ_a \cap \succ_s| &= |(\succ_u \cap \succ_s) \cup (\sim_u \cap \succ_s)| = C^+ + C^s, \\ |\succ_a \cap \succ_s^c| &= |\succ_u \cap \succ_s^c| = C^-, \\ |\succ_a \cap \sim_s| &= |\succ_u \cap \sim_s| = C^u. \end{aligned} \tag{18}$$

From these values, *dpm* is given by:

$$dpm(\succ_u, \succ_s) = \beta(\succ_a, \succ_s) = 2C^- + C^u. \tag{19}$$

Since both C^- and C^u can be directly calculated from $\succ_u$, in practice, it is not necessary to construct the acceptable ranking $\succ_a$. This formula only differs from equation (14), derived from the perfect ranking criterion, by a value of C^s. That is, *dpm* excludes the contribution of these document pairs (d, d') such that $d \sim_u d'$, but $d \succ_s d'$ or $d' \succ_s d$. This result is in fact consistent with the interpretation of the perfect and acceptable ranking criteria.

Example 3 Let

$$\begin{matrix} d_1 \\ d_2 \end{matrix} \succ_u d_3 \succ_u \begin{matrix} d_4 \\ d_5 \end{matrix}$$

be a user ranking on a set of five documents $D = \{d_1, d_2, d_3, d_4, d_5\}$, and

$$\begin{matrix} d_1 \\ d_5 \end{matrix} \succ_s \begin{matrix} d_2 \\ d_4 \end{matrix} \succ_s d_3$$

be a ranking generated by a retrieval system. With respect to $\succ_u$, the closest acceptable ranking to $\succ_s$ is given by:

$$d_1 \succ_a d_2 \succ_a d_3 \succ_a d_5 \succ_a d_4.$$

Based on the performance measure (19), we have:

$$dpm(\succ_u, \succ_s) = \beta(\succ_a, \succ_s) = 8.$$

Note that the same result can be obtained directly from equations (18) and (19), using the user ranking $\succ_u$. □

The performance measure (19) is defined for a single query. In practical situations, it is usually necessary to find the average performance of a system for a group of queries (Cooper, 1968). Let Q be a set of M queries. The mean distance is given simply by:

$$\overline{dpm} = \frac{1}{M}\sum_{i=1}^{M} dpm(\succ_{u_i}, \succ_{s_i}) = \frac{1}{M}\sum_{i=1}^{M} \beta(\succ_{a_i}, \succ_{s_1}) = \frac{1}{M}\sum_{i=1}^{M}(2C_i^- + C_i^u), \tag{20}$$

where $\succ_{u_i}$, $\succ_{s_i}$ and $\succ_{a_i}$ represent, respectively, the user ranking, system ranking and the closest acceptable ranking for a particular query q_i. The motivation for using mean distance measure is similar to that of mean expected search length proposed by Cooper (1968). This means that a retrieval system should be designed to minimize the total distance between a set of user rankings and system rankings.

5 Normalized Performance Measure

The distance-based performance measure *dpm* provides an appropriate basis for comparing various retrieval systems with a fixed query. It may be considered as an absolute (unnormalized) distance function. Although the mean distance measure with absolute distance is appealing, it does not evaluate the performance of every query equally. For example, a performance improvement which reduces distance from 200 down to 100 for one query is considered to be the same as the one which reduces from 2 to 1 for 100 queries. To resolve this problem, relative (normalized) distance measures may be used.

A normalized distance-based performance measure may be defined in terms of distance relative to the maximum distance, namely,

$$ndpm(\succ_u, \succ_s) = \frac{dpm(\succ_u, \succ_s)}{\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ)}, \tag{21}$$

where $\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ)$ is the maximum distance between $\succ_u$ and all rankings. In effect, the actual distance is scaled relative to the potentially realizable distance. The value of *ndpm* lies between 0 and 1. Any acceptable ranking would have a distance of 0, and a ranking farthest away from $\succ_u$ would have a normalized distance of 1. Consequently, all queries are evaluated on a common basis. The use of relative measures can also be found in the comparative percentage improvement figure (Harper & van Rijsbergen, 1978), the normalized recall (Rocchio, 1971) and the expected search length reduction factor (Cooper, 1968).

Intuitively, the worst ranking provided by a retrieval system is the one that reverses the user ranking. Based on the definition of *dpm*, it can be easily verified that the

converse ranking $\succ_u^c$ indeed produces the maximum *dpm* value, namely,

$$\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ) = dpm(\succ_u, \succ_u^c) = 2|\succ_u^c| = 2|\succ_u| = 2C. \tag{22}$$

Moreover, $\succ_u^c$ is the only ranking having the maximum value. Combining equations (19), (21), and (22), the normalized distance-based measure can be calculated by:

$$ndpm(\succ_u, \succ_s) = \frac{dpm(\succ_u, \succ_s)}{dpm(\succ_u, \succ_u^c)} = \frac{2C^- + C^u}{2C}. \tag{23}$$

For example, given the two rankings in Example 3, the normalized measure gives $ndpm(\succ_u, \succ_s) = 8/16$. For a set of M queries, the mean normalized measure can be computed as:

$$\overline{ndpm} = \frac{1}{M}\sum_{i=1}^{M} \frac{dpm(\succ_{u_i}, \succ_{s_i})}{dpm(\succ_{u_i}, \succ_{u_i}^c)} = \frac{1}{M}\sum_{i=1}^{M} \frac{2C_i^- + C_i^u}{2C_i}. \tag{24}$$

The rationale behind the proposed normalized measure is similar to that of the expected search length reduction factor and the expected precision gain factor proposed by Cooper (1968). A system's performance is compared with that of a purely random retrieval. A zero rating will be assigned for a system which is equivalent to a random search of the entire document collection, a positive rating for system which is better than random retrieval, and a negative rating for system which is worse than random retrieval. The purely random retrieval can be characterized by the empty relation $\succ_0 = \emptyset$, which represents the degenerate weak order having only one level. For a user ranking $\succ_u$, it is indeed the acceptable ranking closest to $\succ_0$. There is no contradictory pair and the number of compatible pairs is $|\succ_u|$. Thus,

$$dpm(\succ_u, \succ_0) = \beta(\succ_u, \succ_0) = |\succ_u|. \tag{25}$$

The distance reduction relative to that of the random retrieval, called *distance reduction factor*, is given by:

$$\text{distance reduction factor} \quad = \quad \frac{dpm(\succ_u, \succ_0) - dpm(\succ_u, \succ_s)}{dpm(\succ_u, \succ_0)}$$

$$
\begin{aligned}
&= 1 - \frac{dpm(\succ_u, \succ_s)}{dpm(\succ_u, \succ_0)} \\
&= 1 - \frac{dpm(\succ_u, \succ_s)}{|\succ_u|} \\
&= 1 - 2\, ndpm(\succ_u, \succ_s). \qquad (26)
\end{aligned}
$$

That is, the distance reduction factor is only a transformation of the normalized measure so that its range is $[-1, 1]$, with -1 for the worst performance, 1 for the best performance and 0 for a system equivalent to random retrieval. For the evaluation of system performance, either of these two measures may be used.

6 Relationships of Distance-based Measures to Other Performance Measures

This section shows the inherent relationships between the proposed measure and other standard performance measures.

6.1 Precision, recall and fallout

The standard precision, recall and fallout measures require a two-level relevance judgement, i.e., relevant and non-relevant. According to the results of a retrieval system, a document collection is divided into two subsets, the retrieved subset and non-retrieved subset. Let X denotes the subset of relevance documents and Y the subset of retrieved documents. It is convenient to summarize such information in the following 2×2 table:

	Relevant	Non-relevant	
Retrieved	$X \cap Y$	$\overline{X} \cap Y$	Y
Not retrieved	$X \cap \overline{Y}$	$\overline{X} \cap \overline{Y}$	$\overline{Y}$
	X	$\overline{X}$	D

In this table, $\overline{X} = D - X$ denote the complement of X. The standard precision (P), recall (R) and fallout (F) are defined as:

$$
P = \frac{|X \cap Y|}{|Y|}, \qquad (27)
$$

$$R = \frac{|X \cap Y|}{|X|}, \tag{28}$$

$$F = \frac{|\overline{X} \cap Y|}{|\overline{X}|}. \tag{29}$$

These three measures are related by the following functional relationship:

$$F = \frac{G(1-P)}{(1-G)P}R, \tag{30}$$

where $G = |X|/|D|$ is called generality which measures the density of relevant documents in the collection.

Within this framework, both the relevance judgments and the retrieval results can be expressed in terms of two-level rankings, namely

$$X \succ_u \overline{X} \qquad \text{and} \qquad Y \succ_s \overline{Y}. \tag{31}$$

For these two rankings, we have:

$$\begin{aligned} C^- &= |\succ_u \cap \succ_s^c| \\ &= |(X \times \overline{X}) \cap (\overline{Y} \times Y)| \\ &= |X \cap \overline{Y}||\overline{X} \cap Y|, \end{aligned} \tag{32}$$

$$\begin{aligned} C^u &= |\succ_u \cap \sim_s| \\ &= |(X \times \overline{X}) \cap ((Y \times Y) \cup (\overline{Y} \times \overline{Y}))| \\ &= |X \cap Y||\overline{X} \cap Y| + |X \cap \overline{Y}||\overline{X} \cap \overline{Y}|, \end{aligned} \tag{33}$$

$$C = |\succ_u| = |X \times \overline{X}| = |X||\overline{X}|. \tag{34}$$

According to equation (23), the normalized measure is calculated by:

$$\begin{aligned} ndpm &= \frac{2C^- + C^u}{2C} \\ &= \frac{2|X \cap \overline{Y}||\overline{X} \cap Y| + |X \cap Y||\overline{X} \cap Y| + |X \cap \overline{Y}||\overline{X} \cap \overline{Y}|}{2|X||\overline{X}|} \\ &= \frac{1}{2}\left(1 + \frac{|\overline{X} \cap Y|}{|\overline{X}|} - \frac{|X \cap Y|}{|X|}\right) \\ &= \frac{1}{2}(1 + F - R). \end{aligned} \tag{35}$$

In this special case, the proposed measure may be interpreted as the difference between recall and fallout. From equation (26), the distance reduction factor is defined by $R-F$. This quantity was used by Goffman and Newill (1966; Robertson, 1969) as a measure of retrieval effectiveness. Furthermore, equation (35) can be equivalently expressed as:

$$ndpm = 1 - \frac{1}{2}(R - F + 1) = 1 - A. \tag{36}$$

The quantity $A = (R - F + 1)/2$ is Swets' measure A for the 2×2 table (Robertson, 1969; Swets, 1969).

From equation (30), the proposed measure can also be expressed in terms of precision, recall and generality as:

$$ndpm = \frac{1}{2}\left(1 + \frac{R(G - P)}{P(1 - G)}\right). \tag{37}$$

Therefore, in the case of 2×2 table, *ndpm* may also be viewed a single-valued composite measure of precision and recall. In particular, if P is assumed to be a constant, *ndpm* increases with respect to R for P in $[0, G]$ and decreases for P in $[G, 1]$. If R is assumed to be a constant, *ndpm* decreases with respect to P.

6.2 Marczewski-Steinhaus metric

Heine (1973) suggested that the Marczewski-Steinhaus metric, or MZ-metric for short, may be used to evaluate a retrieval system. For the standard 2×2 table, the MZ-metric is defined by:

$$M = \frac{|X \,\Delta\, Y|}{|X \cup Y|} = 1 - \frac{|X \cap Y|}{|X \cup Y|}, \tag{38}$$

where $X \,\Delta\, Y = (X \cup Y) - (X \cap Y)$ denotes the symmetric difference of two sets. Clearly, the measure M lies between 0 and 1. Based on its properties, the MZ-metric may be viewed as a distance between two sets (Marczewski & Steinhaus, 1958). In terms of precision and recall, the MZ-metric can be expressed by:

$$M = 1 - [P^{-1} + R^{-1} - 1]^{-1}. \tag{39}$$

By combining equations (37) and (39), the following is obtained:

$$ndpm = \frac{1}{2}\left(1 + \frac{(1-M)(G-P)}{(P(2-M)+(M-1))(1-G)}\right). \tag{40}$$

If P is assumed to be a constant, *ndpm* decreases with respect to M for P in $[0, G]$ and increases for P in $[G, 1]$. Similarly, *ndmp* can be expressed through R and M.

The measure M, defined by equation (38), has little in common with *ndpm*. However, when the MZ-metric is applied to $\succ_a$ and $\succ_s$, instead of the sets of relevant and retrieved documents, a very close relationship between MZ-metric and *ndmp* emerges. For two sets (rankings) $\succ_a$ and $\succ_s$, we have:

$$\begin{aligned}
|\succ_u| &= |(\succ_u \cap \succ_s) \cup (\succ_u \cap \succ_s^c) \cup (\succ_u \cap \sim_s)| \\
&= |\succ_u \cap \succ_s| + |\succ_u \cap \succ_s^c| + |\succ_u \cap \sim_s| \\
&= C^+ + C^- + C^u, && (41)\\
|\succ_s| &= |(\succ_u \cap \succ_s) \cup (\succ_u^c \cap \succ_s) \cup (\sim_u \cap \succ_s)| \\
&= |\succ_u \cap \succ_s| + |\succ_u^c \cap \succ_s| + |\sim_u \cap \succ_s| \\
&= C^+ + C^- + C^s, \\
|\succ_a \cap \succ_s| &= |\succ_u \cap \succ_s| + |\sim_u \cap \succ_s| = C^+ + C^s, && (42)\\
|\succ_a \cup \succ_s| &= |\succ_a| + |\succ_s| - |\succ_a \cap \succ_s| \\
&= |\succ_u| + |\sim_u \cap \succ_s| + |\succ_s| - |\succ_a \cap \succ_s| \\
&= 2C^- + C^+ + C^u + C^s, && (43)\\
|\succ_a \Delta \succ_s| &= |\succ_a \cup \succ_s| - |\succ_a \cap \succ_s| \\
&= 2C^- + C^u. && (44)
\end{aligned}$$

Therefore, the application of MZ-metric to $\succ_a$ and $\succ_s$ results in:

$$m = \frac{|\succ_a \Delta \succ_s|}{|\succ_a \cup \succ_s|} = \frac{2C^- + C^u}{2C^- + C^+ + C^u + C^s} = \frac{dpm}{2C^- + C^+ + C^u + C^s}. \tag{45}$$

Comparing this result with that of equation (23), one concludes that the numerator is the unnormalized distance between two ranking. Since $0 \leq m \leq 1$, it provides an alternative normalized distance between two rankings.

Van Rijsbergen (1974, 1979) considered another normalized symmetric distance between two sets, namely:

$$e = \frac{| \succ_a \Delta \succ_s |}{| \succ_u | + | \succ_s |} = \frac{dpm}{| \succ_a | + | \succ_s |}. \tag{46}$$

This measure differs from m only in the normalization denominators. Consider a special case in which $\succ_s$ is a linear order, i.e., there is only one document in each level. If the system ranking $\succ_s$ is a linear order, the closest acceptable ranking $\succ_a$ is also a linear order. In this case,

$$| \succ_a | = | \succ_s | = \frac{1}{2}N(N-1). \tag{47}$$

Substituting it into equation (46), the measure e becomes:

$$e = \frac{dpm}{| \succ_a | + | \succ_s |} = \frac{dpm}{N(N-1)} = \frac{1}{2}(1-\tau), \tag{48}$$

where

$$\tau = 1 - 2\frac{dpm}{N(N-1)}, \tag{49}$$

is the a *correlation coefficient* between two linear orders used by Kendall (1955; Bogart, 1973; Roberts, 1976). In fact, the measure τ is the distance reduction factor for the case of linear orders. For any two arbitrary rankings $\succ$ and $\succ'$, it can be proved that $\beta(\succ, \succ') \leq N(N-1)$. Therefore, equation (48) may be considered as another version of normalized distance in which the normalization factor does not depend on either user or system ranking.

6.3 Normalized recall

In many cases, the standard 2×2 table is only an over-simplification of a real retrieval situation. Consider now a case in which the user ranking is still two levels and the system ranking is a linear order. This situation is referred to as the vertical extension of the standard 2×2 table (Robertson, 1969). For a linear order, Rocchio (1971) proposed the normalized recall measure R_{norm} which can be calculated as follows. A recall versus rank-level graph is drawn for each of the best ranking (i.e., the closest

acceptable ranking), the actual system ranking, and the worst ranking (i.e., the ranking farthest away from $\succ_u$). R_{norm} is computed as the area between the actual case and the worst case relative to the area between the best and worst cases (Bollmann, 1983; Robertson, 1969). Explicitly, R_{norm} is given by:

$$R_{norm} = 1 - \frac{\sum_{i=1}^{n} l_i - \sum_{i=1}^{n} i}{n(N-n)} = 1 - \frac{\sum_{i=1}^{n}(l_i - i)}{n(N-n)}, \tag{50}$$

where $n = |X|$ is the number of relevant documents, $N = |D|$ is the number of all documents in the collection, and l_i is the level of i-th relevant document.

Since there is only one document in each level of the system ranking, the number of compatible pairs is 0. The i-th relevance document on the level l_i induces $(l_i - i)$ contradictory pairs. In sum, for the n relevant documents, the normalized measure gives:

$$ndpm = \frac{2C^-}{2|\succ_u|} = \frac{2\sum_{i=1}^{n}(l_i - i)}{2n(N-n)} = 1 - R_{norm}. \tag{51}$$

Thus, Rocchio's normalized recall may be considered as an *inverse* function of the proposed measure *ndpm*, with 0 for worst performance and 1 for best performance. Robertson (1969) proved that the normalized recall is equivalent to Swets' measure A obtained by the area under the recall-fallout graph on linear scales. Bollmann (1983) pointed out that the normalized recall may be expressed in terms of either the expected search length or the GRE measure used by Noreault, Koll and McGill (1977). This shows that the proposed measure is related to these measures.

If a system ranking is not a linear order but a weak order, it is difficult to interpret and calculate Rocchio's normalized recall. In addition, the normalized recall cannot be directly applied to situations where the user ranking has more than two levels. The solutions proposed by Rocchio (1971) are not very convincing because *ad hoc* methods are involved. To resolve these problems, Bollmann *et al.* (1986) proposed the following generalized normalized recall:

$$\begin{aligned} R_{norm} &= \frac{1}{2}\left(1 + \frac{\text{the No. of agreeing pairs } - \text{ the No. of contradictory pairs}}{\text{the maximum No. of agreeing pairs}}\right) \\ &= \frac{1}{2}\left(1 + \frac{C^+ - C^-}{|\succ_u|}\right) \end{aligned}$$

$$= \frac{1}{2}\left(1 + \frac{C^+ - C^-}{C}\right), \tag{52}$$

which reduces to the Rocchio's normalized recall in the special case discussed earlier. On the other hand, *ndpm* can be rewritten as:

$$\begin{aligned} ndpm &= \frac{2C^- + C^u}{2C} \\ &= \frac{(C^+ + C^- + C^u) - (C^+ - C^-)}{2C} \\ &= \frac{1}{2} - \frac{C^+ - C^-}{2C} \\ &= 1 - \frac{1}{2}\left(1 + \frac{C^+ - C^-}{C}\right) \qquad (53) \\ &= 1 - R_{norm}. \qquad (54) \end{aligned}$$

The same relationship still holds between proposed measure and the generalized normalized recall.

6.4 Ranking-based precision and recall

The standard notions of precision and recall cannot be directly applied to situations where a user ranking has more than two levels, i.e., a horizontal extension of the standard 2×2 table (Robertson, 1969). By modifying these notions, they are applied to the user and system rankings as suggested by Frei and Schäuble (1991).

Given a user ranking $\succ_u$ and a system ranking $\succ_s$, we may conveniently describe them by the following revised 2×2 table:

	User ranking	Non user ranking	
System ranking	$\succ_u \cap \succ_s$	$\overline{\succ}_u \cap \succ_s$	$\succ_s$
Non system ranking	$\succ_u \cap \overline{\succ}_s$	$\overline{\succ}_u \cap \overline{\succ}_s$	$\overline{\succ}_s$
	$\succ_u$	$\overline{\succ}_u$	$D \times D$

where the complement of a relation is defined by $\overline{\succ} = D \times D \; - \; \succ$. If the perfect ranking criterion is adopted, the ranking-based precision (p) and recall (r) are defined as:

$$p = \frac{|\succ_u \cap \succ_s|}{|\succ_s|}, \tag{55}$$

$$r = \frac{|\succ_u \cap \succ_s|}{|\succ_u|}. \quad (56)$$

That is, ranking-based precision is defined as the proportion of the system ranking actually agreeing with the user ranking, and ranking-based recall as the proportion of the user ranking provided by the system ranking.

For the special case where both user and system rankings have only two levels, the ranking-based precision and recall can be computed as:

$$p = \frac{|\succ_u \cap \succ_s|}{|\succ_s|} = \frac{|X \cap Y||\overline{X} \cap \overline{Y}|}{|Y||\overline{Y}|} = P \cdot P', \quad (57)$$

$$r = \frac{|\succ_u \cap \succ_s|}{|\succ_u|} = \frac{|X \cap Y||\overline{X} \cap \overline{Y}|}{|X||\overline{X}|} = R \cdot R', \quad (58)$$

where

$$P' = \frac{|\overline{X} \cap \overline{Y}|}{|\overline{Y}|}, \quad (59)$$

denotes the precision with respect to non-relevant documents, and

$$R' = \frac{|\overline{X} \cap \overline{Y}|}{|\overline{X}|}, \quad (60)$$

the corresponding recall. It is interesting to note that $P' = 1 - F$ and $R' = 1 - B$, where B is a measure used by Robertson (1969). It can be proved that p is a monotonic increasing function of P and r is a monotonic increasing function of R. Thus, it would be expected that there is a similar behavior for both standard and ranking-based recall-precision graphs.

If the acceptable ranking criterion is used, another version of ranking-based precision and recall is given by:

$$p_a = \frac{|\succ_a \cap \succ_s|}{|\succ_s|}, \quad (61)$$

$$r_a = \frac{|\succ_a \cap \succ_s|}{|\succ_a|}. \quad (62)$$

Using p_a and r_a, performance measures m and e, defined by equations (45) and (46), can now be rewritten as:

$$m = \frac{|\succ_a \Delta \succ_s|}{|\succ_a \cup \succ_s|} = 1 - [p_a^{-1} + r_a^{-1} - 1]^{-1}, \quad (63)$$

$$e = \frac{| \succ_a \Delta \succ_s |}{| \succ_a | + | \succ_s |} = 1 - 2[(p_a^{-1} + r_a^{-1}]^{-1}. \quad (64)$$

The proposed measure *ndpm* may be expressed as:

$$ndpm = \left(1 - \frac{2}{p_a^{-1} + r_a^{-1}}\right) \frac{| \succ_a | + | \succ_s |}{2| \succ_u |}. \quad (65)$$

Thus, the distance-based measures may be interpreted as composite measures of ranking-based precision and recall.

To some extent, the appropriateness of the proposed distance-based performance measures depends on whether axioms 1-4 are meaningful empirically. Although one may question the validity of these axioms, they do clearly state the conditions under which one may use a distance-based measure. The close relationships between the proposed measures and many existing measures examined in this section suggest that similar axioms are indeed implicitly adopted. The explicit statement of the axioms involved makes the proposed measures to be more transparent, on which further investigation may be based.

7 Conclusion and Futher Research

Based on the notion of user preference, an attempt is made to examine the fundamental issues regarding the representation, interpretation, and measurement of user judgments on documents. It seems reasonable that a user preference relation on documents must obey two basic axioms, asymmetry and negative transitivity. This guarantees the measurement of user judgments with an ordinal scale. The use of an ordinal scale implies one must be cautious when using the absolute relevance value of a document. A performance measure that uses the information about the relative order induced by the relevance values confirms to the valid use of an ordinal scale, as it is invariant to strictly monotonic increasing transformations of the absolute relevance values. A performance measure using such information is proposed based on a distance function between user and system rankings. The distance function is uniquely determined by a set of reasonable axioms. In special cases, the proposed measure has

close relationships with the standard measures, such as recall, precision, normalized recall, and MZ-metric.

There are two types of uesr preference in information retrieval, the user preference on documents and the user preference on retrieval results (i.e., system rankings). By examining the user preference on retrieval results, many authors have attempted to establish a basis for existing performance measures (Bollmann, 1977; Bollmann & Cherniavsky, 1981; Bollmann & Raghavan, 1988; Cherniavsky & Lakhuty, 1971; van Rijsbergen, 1974). Their investigations consider a simple two-level user preference structure on documents. On the other hand, this paper only concentrates on the user preference on documents. The use preference structure on system rankings is not explicitly stated. The discussion relies on an implicit, and intuitively appealing, assumption that a user would prefer a ranking closer to the user ranking to another ranking further away. The closeness between rankings is measured by a distance function. The choice of the distance function is justified by quantitative axioms 1-4. Moreover, the use of arithmetic mean distance suggests that it should be based on an interval scale.

The present investigation and the existing studies are complementary to each other. Each of them captures some important but distinct aspects of system evaluation by focusing on different types of user preference. A measurement-theoretic foundation of system evaluation should take both types of preference into account. It is important to establish a more general framework by extending and combining these results. In particular, we need to examine the empirical validity of axioms 1-4. We also need to investigate the qualitative properties of the user preference on retrieval results that permit the use of the proposed performance measures as an interval scale. The results of such further research may lead to a more solid measurement-theoretic foundation for system evaluation.

Many practical issues have not been addressed in this study. The application of the proposed measure requires the user preference over the entire document collection that is usually not available. It is important to ascertain how to use this measure

when only partial user preference information is given (Frei & Schäuble, 1991; Fuhr, 1989). An empirical examination of the measure will be useful and complement to the theoretical analysis.

Acknowledgements

The author is grateful for financial support from NSERC Canada, editorial help from E.H. Dale, discussion with and suggestions from P. Bollmann about the normalized recall, and for the constructive comments from the anonymous referees.

References

Bogart, K.P. (1973). Preference structures I: distance between transitive preference relations. *Journal of Mathematical Sociology, 3*, 49-67.

Bollmann, P. (1977). A comparison of evaluation measures for document retrieval systems. *Journal of Informatics, 1*, 97-116.

Bollmann, P. (1983). The normalized recall and related measures. Unpublished manuscript.

Bollmann, P., & Cherniavsky, V.S. (1981). Measurement-theoretical investigation of the MZ-metric. In Oddy, R.N., Robertson, S.E., & van Rijsbergen, C.J. (Eds.), *Information Retrieval Research*. London: Butterworths.

Bollmann, P., Jochum, R., Reiner, U., Weissmann, V., & Zuse, H. (1986). Planung und durchführung der retrievaltests. In Schneider, H. (Ed.), *Leistungsbewertung von Information Retrieval Verfahren (LIVE)*, 183-212.

Bollmann, P., & Raghavan, V.V. (1988). A utility-theoretic analysis of expected search length. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Bollmann, P., Raghavan, V.V., Jung, G.S., & Shu, L.C. (1992). On probabilistic notions of precision as a function of recall. *Information Processing and Mangement, 28*, 291-315.

Bollmann, P., & Wong, S.K.M. (1987). Adaptive linear information retrieval models. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 157-163.

Bookstein, A. (1979). Relevance. *Journal of the American Society for Information Science, 30*, 269-273.

Bookstein, A. (1983). Outline of a general probabilistic retrieval model. *Journal of Documentation, 39*, 63-72.

Bookstein, A. (1989). Set-oriented retrieval. *Information Processing and Management, 25*, 465-475.

Cherniavsky, V.S., & Lakhuty, D.G. (1971). On the problem of information system evaluating. *Automatic Documentation and Mathematical Linguistics, 4*, 9-26.

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. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*, 19-37.

Cuadra, C.A., & Katter, R.V. (1967). *Experimental Studies of Relevance Judgments.* TM-3520, Systems Development Corp., Santa Monica, California.

Eisenberg, M. (1988). Measuring relevance judgments. *Information Processing and Management, 24*, 373-389.

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

Fishburn, F.C. (1970). *Utility Theory for Decision Making.* New York: Wiley.

Frei, H.P., & Schäuble, P. (1991). Determining the effectiveness of retrieval algorithms. *Information Processing and Mangement, 27*, 153-164.

French, S. (1986). *Decision Theory – An Introduction to the Mathematics of Rationality.* Chichester: Ellis Horwood Limited.

Fuhr,, N. (1989). Optimum polynomial retrieval functions based on probability ranking principle. *ACM Transactions on Information System, 3*, 183-204.

Goffman, W., & Newill, V.A. (1966). A methodology for test and evaluation of information retrieval systems. *Information Retrieval and Storage, 3*, 19-25.

Harper, D.J., & van Rijsbergen, C.J. (1977). An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation, 34*, 189-216.

Heine, M.H. (1973). Distance between sets as an objective measure of retrieval effectiveness. *Information Retrieval and Storage, 9*, 181-198.

Janes, J. W. (1991). The binary nature of continuous relevance judgments: a study of users' perceptions. *Journal of the American Society for Information Science, 42*, 754-756.

Keen, E.M. (1971). Evaluation parameters. In Salton, G. (Ed.), *The SMART Retrieval System — Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall, 74-111.

Kemeny, J.G., & Snell, J.L. (1962). *Mathematical Models in the Social Sciences*. New York: Blaisdell.

Kendall, M.G. (1955). *Rank Correlation Methods*. New York: Hafner Publishing Company.

King, D.W. (1968). Design and evaluation of information systems. *Annual Review of Information Science and Technology, 3*, 61-103.

Kochen, M. (1974). *Principles of Information Retrieval*. Los Angeles: Nelville.

Lesk, M.E., & Salton, G. (1968). Relevance accessments and retrieval system evaluation. *Information Retrieval and Storage, 4*, 317-323.

Marczewski, E., & Steinhaus, H. (1958). On a certain distance of sets and the corresponding distance of functions. *Colloquium Mathemmaticum, 6*, 319-327.

Maron, M.E., & Kuhns, J.L. (1970). On relevance, probabilistic indexing and information retrieval. In Saracevic, T. (Ed.), *Introduction to Information Science*. New York: R.R. Bowker Company, 295-311.

Noreault, T., Koll, M., & McGill, M.J. (1977). Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science, 28*, 333-339.

Raghavan, V.V., Bollmann, P., & Jung, G.S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information System, 7*, 205-229.

Rees, A.M., & Schultz, D.G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching. *Final Report to the National Science Foundation.* Center for Documentation and Communication Research, Case Western Reserve University, Cleveland, Ohio.

Roberts, F.S. (1976). *Discrete Mathematical Models.* Englewood Cliffs, New Jersey: Prentice-Hall.

Roberts, F.S. (1979). *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Reading, Massachusetts: Addison-Wesley.

Robertson, S.E. (1969). The parametric description of retrieval tests, part I: the basic parameters, part II: overall measures. *Journal of documentation, 25*, 1-27, 93-107.

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

Rocchio, Jr. J.J. (1971). Performance indices for document retrieval. In Salton, G. (Ed.), *The SMART Retrieval System — Experiments in Automatic Document Processing.* Englewood Cliffs, NJ: Prentice-Hall, 57-67.

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing and Management, 28*, 441-449.

Salton, G. & McGill, M.H. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Saracevic, T. (1970). The concept of "relevance" in information science: a historical review. In Saracevic, T. (Ed.), *Introduction to Information Science.* New York: R.R. Bowker Company, 111-151.

Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*, 321-343.

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivison, D. (1987). A study of information seeking and retrieving, I: background and methodology. *Journal of the American Society for Information Science, 39*, 161-176.

Sparck Jones, K. (Ed.) (1981). *Information Retrieval Experiment*. London: Butterworths.

van Rijsbergen, C.J. (1974). Foundation of evaluation. *Journal of Documentation, 30*, 365-373.

van Rijsbergen, C.J. (1979). *Information Retrieval*. London: Butterworths.

Swets, J.A. (1969). Effectiveness of information retrieval methods. *American Documentation, 20*, 72-89.

Wong, S.K.M., Bollmann, P., & Yao, Y.Y. (1991). Information retrieval based on axiomatic decision theory. *International Journal of General Systems, 19*, 107-117.

Wong, S.K.M., & Yao, Y.Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Science, 41*, 334-341.

Wong, S.K.M., Yao, Y.Y., & Bollmann, P. (1988). Linear structure in information retrieval. *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 219-232.