1980 TOC Search

16

Measurement-theoretical investigation of the MZ-metric

P. Bollmann and V. S. Cherniavsky

16.1 Introduction

Information retrieval has a long and rich tradition of measuring and evaluating. Many measures have been suggested for evaluating performance and quality of information retrieval systems. There are many examples where large-scale experiments have been carried out aiming at evaluating and comparing different systems. At the same time, it is well known that different formal measures reflect different intuitive ideas of quality and have different properties. They often contradict each other and they are often incompatible. However, there is as yet no systematic way to describe and investigate the relation between intuitive ideas of quality, on the one hand, and measures representing formally those ideas of quality, on the other. Without a clear understanding of this relation there is no sound basis for comparing different measures or for selecting in some given applicational environment one of the many measures suggested to date. Having no justifiable criteria for comparing different measures and for selecting one of them, users apply formal criteria such as the measure having one number as value, or having a maximum or minimum (Swets, 1969). Such decisions, as well as judgements about relative quality of information retrieval systems based on those decisions, are not convincing: one system cannot be declared to be better than another one just because some formal mechanism assigns to the first system a symbol which in a formal ordering lies higher than the symbol assigned to the second system. Information systems are conceived for practical use and their quality is a practical matter rather than a formal one.

This situation has given and still gives rise to protests. Users try to overcome it in two ways. The first way leads to requiring the measure applied to have certain measurement-theoretic properties (van Rijsbergen, 1974). The second way tries to justify measures by connecting them to the practical application of the information system evaluated by means of this measure (Lancaster, 1968). Both ideas seem sound. However, leaning upon measurement theory, without unambiguous emphasis on the intuitive viewpoint about quality, amounts to justifying one formalism by another. This way the problem of justification is postponed rather than solved. The other idea, trying to connect measures with practical applications, of the system measured, would be precisely what is needed if we could achieve it. However, it does not work so far. Therefore, it seems reasonable to divide this step into two steps. The intuitive idea of quality (of the system's output) is the link connecting the practical application of the information system, on the one hand, with the formal measure evaluating the performance of this system, on the other. Hence, developing a mechanism for describing and investigating the relation between formal measures for information systems' performance, on the one hand, and the intuitive ideas, or viewpoints, about the quality of those systems, on the other, seems to be a step towards the attempted goal.

Such a mechanism is the central point of this chapter. This mechanism has been developed in Cherniavsky and Lakhuty (1970) and Bollmann and Cherniavsky (1977, 1979), in the form of what we call there *transformational viewpoints.* These transformational viewpoints are relations connecting pairs of information retrieval systems' outputs. In this chapter, transformational viewpoints are called *elementary viewpoints.* Such an elementary viewpoint is a set of output pairs, each pair being interpreted as 'the first element of the pair is not worse than the second one'. Hence, an elementary viewpoint is, in fact, a formalised viewpoint on quality of the information systems' outputs. Moreover, we indicate methods for constructing combined viewpoints from elementary ones. A combined viewpoint, described in terms of elementary viewpoints, is thereby formalised to the extent necessary for precisely comparing it with formal properties of performance measures. The set of combined viewpoints we are able to construct in terms of elementary viewpoints is fairly large. In fact, in this way we are able to indicate the intuitive viewpoints on quality for most popular formal measures known to date.

There is also another problem connected with the application of performance measures. As a matter of fact, all measures suggested to date evaluate outputs of systems rather than the systems themselves. In order to obtain an evaluation of a system, we have to consider it as being represented by the set of its outputs. Applying a measure, we have then a set of measurement values. In order to transform it into an evaluation value for the ysstem, we have to apply some statistics — for example, the arithmetic mean, the median, etc. However, the application of a statistic is limited by the measurement-theoretic properties of the initial measure. These properties are not just formal properties of the formal measure. They are, rather, additional assumptions made on the basis of the intuitive viewpoint about quality which we connect with the measure. Hence, in order to examine the possibility of applying a statistic, which is necessary in order to obtain a system evaluation from a set of output evaluations, we have to consider the pair-measure viewpoint rather than the measure itself. The mechanism of elementary viewpoints is the only mechanism known to us allowing of this kind of consideration.

In this chapter, we apply the mechanism of elementary viewpoints to the so-called MZ-metric (Heine, 1973) considered as a performance measure. This choice is justified by the fact that the MZ-metric is, to the best of our knowledge, the only performance measure which has given rise to a measurement-theoretic discussion and for which a number of stated measurement-theoretic problems exist (van Rijsbergen, 1974, 1978). Correspondingly, our aim in this chapter is twofold: we aim at investigating the MZ-metric and discussing the measurement-theoretic problems attached to it; at the same time, we want to illustrate the application of the mechanism of elementary viewpoints. Space limitations make complete proofs impossible. The interested reader is referred to Bollmann and Cherniavsky (1980).

After this introduction the chapter is structured as follows.

16.2 The *MZ*-metric: Definition. Rijsbergen's investigations. Open questions and problems.
16.3 Some measurement–theoretic notions: What is measuring in general. Scales. Bisymmetric structures. Distributions as retrieval systems' outputs. Viewpoints as intuitive ideas of quality.
16.4 Elementary viewpoints and transformations: Elementary viewpoints. Transformations. Describing measures and viewpoints by elementary viewpoints. *MZ*-metric as ordinal scale. Describing improvements by transformations. Necessary and sufficient conditions for the *MZ*-metric to be an interval scale.
16.5 Discussion of some problems: Representing the *MZ*-metric by the additive conjoint measurement. Replacing recall-precision by recall-fallout. More than two factors.
16.6 Conclusion.

16.2 The *MZ*-metric

The *MZ*-metric and some measures related to it have been studied and discussed by many authors, especially by Heine (1973) and van Rijsbergen (1974, 1979). In 1974 Rijsbergen had been the first to study these measures from a measurement–theoretic point of view. The technique he applied in this investigation was the additive conjoint measurement. It is important that he takes as set of objects to be measured the set of pairs $\{(R, P)\}$, where $0 \leqslant R \leqslant 1$ means recall and $0 \leqslant P \leqslant 1$ means precision. He defines the *MZ*-metric E as a function

$$E\colon [0, 1] \times [0, 1] \rightarrow [0, 1]$$

by the formula

$$E(R, P) = 1 - \frac{1}{\frac{1}{R} + \frac{1}{P} - 1}$$

However, this model is not complete. Indeed, both recall and precision are themselves measurement results rather than measured objects. The original objects we want to measure and to evaluate are the outputs of some information retrieval system. Therefore, in order to make the model complete, we introduce explicitly these outputs in our model and describe them by contingency tables. We define a contingency table to be a quadruple (a, b, c, d) with

a = number of relevant retrieved documents;
b = number of non-relevant retrieved documents;
c = number of relevant not retrieved documents;
d = number of non-relevant not retrieved documents.

We define the MZ-metric E, as applied to contingency tables, by the formula

$$E(a, b, c, d,) = 1 - \frac{a}{a+b+c}$$

where $a+b+c \geqslant 1$ is assumed to hold. In the case in which we assume the document collection to contain at least one relevant document, the above restriction holds always.

With this model the following problems are connected, most of which are formulated by van Rijsbergen (1974, 1979).

(1) Assume that we have a measure applied to (R, P) pairs as objects to be evaluated. What is the effect of replacing the set of (R, P) pairs by the set of contingency tables? In particular, what is the effect of this replacement in the case where the measure is the MZ-metric (van Rijsbergen, 1979)? How are the axioms for additive conjoint measurement affected by the replacement mentioned above?

(2) In the case of the additive conjoint measurement, is it possible to replace the measures recall and precision by the measures recall and fallout (van Rijsbergen, 1974)? What can be the advantages of this replacement (van Rijsbergen, 1979)?

(3) The model of the additive conjoint measurement yields criteria for the component measures to be interval scales. However, it does not yield any criteria for the resulting measure to be an interval scale. What are the corresponding criteria?

(4) Is it possible to extend the additive conjoint measurement to more than two factors (van Rijsbergen, 1974)?

(5) Rijsbergen describes axiomatically a set of measures encompassing the MZ-metric and some other related measures (van Rijsbergen, 1974). Is it possible to define a smaller set still encompassing the above measures?

In this chapter we intend to analyse these questions.

16.3 Some measurement-theoretic notions

Some measurement–theoretic notions are introduced which we use in the subsequent discussion. In the definitions we follow Krantz *et al.* (1971) and Pfanzagl (1973). We specify these notions according to our needs.

Definition: Let A be a set of objects and $R_1, \ldots, R_i, \ldots, R_m$ be a set of n_i-ary relations over A. The tuple $\mathbb{A} = (A, R_1, \ldots, R_m)$ is called a *relational system*.

Definition: Let $\mathbb{A} = (A, R_1, \ldots, R_m)$ be an empirical relational system and $\mathbb{B} = (\mathbb{R}, S_1, \ldots, S_m)$ be a numerical relational system ($\mathbb{R}$ is the set of real numbers). Further, let μ: $A \rightarrow \mathbb{R}$ be a function such that $R_i(a_i, \ldots, a_{n_i})$ holds if and only if $S_i(\mu(a_1), \ldots, \mu(a_{n_i}))$ holds. Then the triple $(\mathbb{A}, \mu, \mathbb{B})$ is called a *one-dimensional real scale*.

Given $\mathbb{A}$ and $\mathbb{B}$, there is usually more than one mapping μ such that $(\mathbb{A}, \mu, \mathbb{B})$ is a scale.

Definition: Consider some scale $(\mathbb{A}, \mu, \mathbb{B})$. This scale is said to be an *ordinal scale* if and only if for each monotonic transformation μ' of μ the triple $(\mathbb{A}, \mu', \mathbb{B})$ is a scale also. In the case for each positive linear transformation $\mu' = \alpha\mu + \beta(\alpha > 0)$ of μ the triple $(\mathbb{A}, \mu', \mathbb{B})$ is a scale, the initial scale $(\mathbb{A}, \mu, \mathbb{B})$ is said to be an *interval scale*.

Definition: Let A be a set and $\cdot\!\geqslant$ a binary relation on A. The relational system $(A, \cdot\!\geqslant)$ is said to be a *weak order* if and only if for all a, b, $c \varepsilon A$ the following always hold:

(i) $a \cdot\!\geqslant b$ or $b \cdot\!\geqslant a$
(ii) $a \cdot\!\geqslant b$, $b \cdot\!\geqslant c$ implies $a \cdot\!\geqslant c$

In the case where $a \cdot\!\geqslant b$ and $b \cdot\!\geqslant a$ hold, we write $a \sim b$ and call a and b *equivalent*. In the case where $a \cdot\!\geqslant b$ and $\urcorner(a \sim b)$ hold we write $a > b$.

Definition: Let A be a non-empty set and $\cdot\!\geqslant$ a binary relation on A. Let $\cdot$ be a binary operation of the type $A \times A \rightarrow A$. We introduce a ternary relation $\cdot(a_1, a_2) = a_3$. For simplicity we use for this relation the same symbol $\cdot$, the number of arguments being sufficient for avoiding ambiguities. Hence, we write $\cdot(a_1, a_2, a_3)$ if and only if $a_1 \cdot a_2 = a_3$. The relational system $(A, \cdot\!\geqslant, \cdot)$ is called a *bisymmetric structure* if and only if for all a, b, b', b'', c, $d \in A$ the following five axioms are fulfilled:

(1) $(A, \cdot\!\geqslant)$ is a weak order.
(2) "$a \cdot\!\geqslant b$" $\equiv$ "$a{\cdot}c \cdot\!\geqslant b{\cdot}c$" $\equiv$ "$c{\cdot}a \cdot\!\geqslant c{\cdot}b$".
(3) $(a \cdot b) \cdot (c \cdot d) = (a \cdot c) \cdot (b \cdot d)$ bisymmetry).
(4) If $b'{\cdot}c \cdot\!\geqslant a \cdot\!\geqslant b''{\cdot}c$ hold, then there exists $\tilde{b}$, such that $\tilde{b}{\cdot}c \sim a$ holds. If $c{\cdot}b' \cdot\!\geqslant a \cdot\!\geqslant c{\cdot}b''$ holds, then there exists $\tilde{\bar{b}}$, such that $c{\cdot}\tilde{\bar{b}} \sim a$ holds. (Restricted solvability.)
(5) Every strictly bounded standard sequence is finite. Here a sequence $\{a_i | a_i \in A, i \in N\}$ is called a *standard sequence* if and only if there exist $p, q \in A$ such that $p > q$ and for all $i \in N$, $a_i \cdot p \sim a_{i+1} \cdot q$ or $p \cdot a_i \sim q \cdot a_{i+1}$.

Theorem 16.1: Let $\boldsymbol{A} = (A, \geqslant, \cdot)$ be a bisymmetric structure. Then there exists a bisymmetric structure $\boldsymbol{B} = (\mathfrak{R}, \geqslant, *)$ with $x * y = \alpha x + \beta y + \lambda$ holding for all x, $y \in \mathfrak{R}$. Here $\alpha > 0$, $\beta > 0$, λ are real numbers. Furthermore, there exists a function φ: $A \rightarrow \mathfrak{R}$ with the following properties:

(1) $x \cdot\!\geqslant y$ if and only if $\varphi(x) \geqslant \varphi(y)$;
(2) $\varphi(x \cdot y) = \varphi(x) * \varphi(y)$
(3) $(\boldsymbol{A}, \varphi, \boldsymbol{B})$ is an interval scale.

We need to specify the above measurement–theoretic notions, adjusting them to the needs of information retrieval. We first define the objects which we measure.

We accept as objects to be measured the answers of a retrieval system to a request. We assume that any request specifies every document in the document collection as being relevant or non-relevant. In this situation, any two relevant documents are considered equivalent, as also are any two non-relevant ones. Therefore no information is lost if every relevant document is represented by a '+' sign and every non-relevant document is represented by a '−' sign. We take an information system's answer to be a partition of the document collection into two sets: the retrieved ones and the non-retrieved ones. Hence, the objects to be measured are ordered pairs of sets called 'ranks' containing each a number of + and − signs. We call these pairs *distributions*. *Figure 16.1* gives an example of a distribution. We assume the left-hand set to be the set of retrieved documents. Correspondingly, the right-hand set is the set of the not-retrieved ones. In a natural way, each distribution defines a contingency table as described in Section 16.2. For the example of *Figure 16.1* the corresponding contingency table is the following: (3, 1, 2, 4). We limit the set $\mathscr{D}$ of objects to

$$\left(\begin{array}{c|c} +\,+\,+ & +\,+ \\ - & -\,-\,-\,- \end{array}\right)$$

Figure 16.1 A distribution

be measured to the set of all distributions Δ such that $a+b+c \geqslant 1$, where a is the number of relevant retrieved documents in Δ, b is the number of non-relevant retrieved documents in Δ, and c is the number of relevant non-retrieved documents in Δ. We shall use the notions 'distribution' and 'contingency table' synonymously, as long as this does not lead to ambiguities.

An evaluation measure φ is a mapping φ: $\mathscr{D} \rightarrow \mathbb{R}$ of the set $\mathscr{D}$ of distributions into the set of real numbers.

In order to evaluate and compare distributions, an intuitive idea is needed telling what it means for a distribution to be good or to be better than another one. This idea of quality cannot be of a formal nature. It must come from outside of the formal evaluation environment and it always depends on specific practical circumstances. This intuitive idea of quality is what makes the evaluation and comparison procedure meaningful. The notion of the intuitive idea of quality has been introduced under the name 'viewpoint' in Cherniavsky and Lakhuty (1970). Hence, the theory of measurement in information retrieval is based on the assumption that some viewpoint is given. This corresponds to the fundamental assumption of general measurement theory that some empirical relational system is given. A viewpoint considered from inside of the evaluation environment is a preference relation $\cdot\!\geqslant$, such that $\Delta\cdot\!\geqslant \Delta'$ is interpreted as "Δ is not worse than Δ'" (from the adopted viewpoint). This may also be interpreted in the sense that the assumed user is willing to give Δ' for Δ. For the preference relations the following hold:

(i) $\Delta \cdot\!\geqslant \Delta$ (reflexivity)
(ii) $\Delta \cdot\!\geqslant \Delta' \Lambda\ \Delta' \cdot\!\geqslant \Delta''$ implies $\Delta \cdot\!\geqslant \Delta''$ (transitivity)

If, in addition, for any Δ and Δ' the following holds

$\Delta \cdot\!\geqslant \Delta'$ or $\Delta' \cdot\!\geqslant \Delta$

then the viewpoint is a weak order as defined above. In a natural way, a viewpoint is thought of and treated as a set $s \subset \mathscr{D} \times \mathscr{D}$.

Let φ: $\mathscr{D} \rightarrow \mathbb{R}$ be an evaluation measure and $\cdot\!\geqslant$ be a viewpoint. In the case where $\Delta \cdot\!\geqslant \Delta'$ if and only if $\varphi(\Delta) \geqslant \varphi(\Delta')$, $\cdot\!\geqslant$ is called the *maximal viewpoint* for φ and denoted by $\cdot\!\geqslant_{\varphi}$.

Let φ: $\mathscr{D} \rightarrow \mathbb{R}$ be an evaluation measure and $D=(\mathscr{D}, \cdot\!\geqslant_{\varphi})$ and $B=(\mathbb{R}, \geqslant)$ be two relational systems. Then the triple (D, φ, B) is a real one-dimensional scale. It is, moreover, an ordinal scale.

16.4 Elementary viewpoints and transformations

As has already been indicated, a viewpoint is, formally speaking, a weak order. However, it would be hard to work with a viewpoint defined in such a general way. In order to have an effective mechanism for describing and analysing viewpoints, we proceed as follows.

We first introduce the notion of the *elementary viewpoint*. An elementary viewpoint V is a binary relation $V \subset \mathscr{D} \times \mathscr{D}$ which, on the one hand, is effective and intuitively clear, so that we can really work with it, and, on the other hand, allows us to describe viewpoints in a sense to be defined below. In this chapter we call the *basis of an elementary viewpoint* a pair (τ, χ) where τ is a mapping τ: $\mathscr{D} \rightarrow \mathscr{D}$ and χ is a three-valued function χ: $\mathscr{D} \rightarrow$ {'better', 'worse', 'equivalent'}. We call τ a *transformation* and χ *a characteristic function.* We require τ and χ to be effective in some sense (for example, in the sense of being algorithmically computable) and to be intuitively clear. An elementary viewpoint $V \subset \mathscr{D} \times \mathscr{D}$ is constructed from its base (τ, χ) in the following way:

(i) $(\Delta, \tau(\Delta)) \varepsilon V$ if and only if $\chi(\Delta)$ = 'worse' or $\chi(\Delta)$ = 'equivalent'
(ii) $(\tau(\Delta), \Delta) \varepsilon V$ if and only if $\chi(\Delta)$ = 'better' or $\chi(\Delta)$ = 'equivalent'
(iii) $(\Delta, \Delta') \varepsilon V$ if and only if it follows from (i) or (ii)

As usual, we write $\Delta \geqslant \Delta'$ for $(\Delta, \Delta') \varepsilon V$.

Let $\mathscr{M} = \{M_i\}_{i \in I}$ be a set of binary relations $M_i \subset \mathscr{D} \times \mathscr{D}$ and let $s \subset \mathscr{D} \times \mathscr{D}$ be a viewpoint. In the case where

$$s = H\left(\bigcup_{i \in I} M_i\right)$$

holds, (here $H(M)$ is the reflexive and transitive closure of $M \subset \mathscr{D} \times \mathscr{D}$) we say that s *is generated by* $\mathscr{M}$. In the sense of this definition, we will describe viewpoints s by sets of elementary viewpoints $\mathscr{V} = \{V_i\}_{i \in I}$.

The following theorem holds:

Theorem 16.2: Let μ_1: $\mathscr{D} \rightarrow \mathfrak{R}$ and μ_2: $\mathscr{D} \rightarrow \mathfrak{R}$ be two measures with the maximal viewpoints s_1 and s_2, respectively. Further, let $\mathscr{V} = \{V_i\}_{i \in I}$ be a set of elementary viewpoints generating s_1. Further, let (τ_i, χ_i), $i \in I$ be the bases of the elementary viewpoints V_i, $i \in I$ respectively. Finally, let

$(\Delta, \tau_i(\Delta)) \in s_1$ if and only if $(\Delta, \tau_i(\Delta)) \in s_2$

and

$(\tau_i(\Delta), \Delta) \in s_1$, if and only if $(\tau_i(\Delta), \Delta) \in s_2$

hold for all $\Delta \in \mathscr{D}$. Then $s_1 = s_2$.

It is now easily seen that under the conditions of theorem 16.2 the measures μ_1 and μ_2 are either equal or monotonic transformations of each other.

We want now to define a set of elementary viewpoints which generate the maximal viewpoint of the MZ-metric E, thus describing this measure. To this end we first define three transformations τ_1, τ_2, τ_3 and an infinite set of transformations $\{\tau_4^k\}_{k=2,3,...}$. We then compare each of these transformations with the measure E in order to specify a characteristic function for each transformation. In this way we obtain the required elementary viewpoints. What then remains is to show that they really generate the maximal viewpoint of E.

(1) The transformation τ_1 is defined as follows: 'Replace a not-retrieved relevant document by a retrieved relevant one, if possible. Otherwise leave the distribution unchanged.' This transformation is equivalent to moving one + sign from the second rank into the first one. This transformation yields the following mapping:

$$\tau_1(a, b, c, d) = \begin{cases} (a+1, b, c-1, d) & c>0 \\ (a, b, c, d) & c=0 \end{cases}$$

Comparing τ_1 with the maximal viewpoint of E yields

$$\tau_1(\Delta) \begin{cases} \cdot \geqslant \Delta, & c>0 \\ \sim \Delta, & c=0 \end{cases}$$

Denoting by V_1 the first elementary viewpoint, we have $(\Delta, \Delta') \in V_1$ if and only if $\Delta = \tau_1(\Delta')$.

(2) The transformation τ_2 is defined by: 'Replace, if possible, one retrieved non-relevant document by one not-retrieved relevant document. Else leave the distribution unchanged.' This yields the following mapping:

$$\tau_2(l, b, c, d) = \begin{cases} (a, b-1, c+1, d) & b>0 \\ (a, b, c, d) & b=0 \end{cases}$$

The comparison with E shows $\tau_2(\Delta) \sim \Delta$. For the elementary viewpoint V_2 we obtain $(\Delta, \Delta') \in V_2$ if and only if $\Delta = \tau_2(\Delta')$ or $\Delta' = \tau_2(\Delta)$.

(3) The transformation τ_3 is defined by: 'Remove, if possible, one not-retrieved non-relevant document from the distribution. Else leave the distribution unchanged.' This yields the mapping

$$\tau_3(a, b, c, d) = \begin{cases} (a, b, c, d-1) & d>0 \\ (a, b, c, d) & d=0 \end{cases}$$

The comparison with E shows $\tau_3(\Delta) \sim \Delta$. For the elementary viewpoint V_3 we obtain $(\Delta, \Delta') \in V_3$ if and only if $\Delta = \tau_3(\Delta')$ or $\Delta' = \tau_3(\Delta)$.

(4) The transformation τ_4^k, $k \geqslant 2$ is defined by: 'Replace each document by k equivalent ones.' This yields the mapping

$$\tau_4^k(a, b, c, d) = (ka, kb, kc, kd)$$

The comparison with E shows $\tau_4^k(\Delta) \sim \Delta$. For the elementary viewpoint V_4^k we obtain $(\Delta, \Delta') \in V_4^k$ if and only if $\Delta = \tau_4^k(\Delta')$ or $\Delta' = \tau_4^k(\Delta)$.

Theorem 16.3: Denoting by $s_E \subset \mathscr{D} \times \mathscr{D}$ the maximal viewpoint of E, the following holds:

$$H\left(V_1 \cup V_2 \cup V_3 \cup \bigcup_{k=2}^{\infty} V_4^k \right) = s_E$$

The inclusion of the left-hand side into the right-hand side has already been shown in comparing the transformations τ_1, τ_2, τ_3, τ_2^k with E. The converse inclusion is proved by referring to the fact that the left-hand side is a transitive closure.

According to the remark to theorem 16.2, the measure E is either equal to or a monotonic transformation of every other measure having the same relation to transformations τ_1, τ_2, τ_3, τ_4^k as measure E has.

Thus, consider the measure F_1,

$$F_1(a, b, c, d) = 1 - \frac{2a}{2a+b+c}$$

of Jardine and Rijsbergen (1971), and the measure F_2,

$$F_2(a, b, c, d) = 1 - \frac{a}{a + 2(b + c)}$$

of Vickery (van Rijsbergen, 1974). It may easily be shown that the above comments apply to F_1 as well as to F_2. This implies that the measures E, F_1 and F_2 are monotonic transformations of one another or, in other words, that they are all equivalent as ordinal scales. The same obviously holds for every measure of the form

$$F_{\alpha,\beta}(a, b, c, d) = 1 - \frac{\alpha a}{\alpha a + \beta(b + c)} \qquad \alpha, \beta > 0$$

which is a known fact in the theory of distance measures (Steinhausen and Langer, 1977).

So far we have discussed the use of transformations and elementary viewpoints for the description of measures used as ordinal scales. However, it seems natural to try to use the same mechanism to analyse measures as defining interval scales. This idea may be realised and we will show this again using the example of the MZ-metric E.

Our plan will be as follows. Measure E induces a measure E^* defined on pairs of distributions by $E^*(\Delta, \Delta') = E(\Delta') - E(\Delta)$. This, in turn, implies a formal notion of improvement which may be used when it is in conformity with the intuitive notion of improvement underlying the application. This is exactly what we mean when we speak about the possibility of using E as an interval scale. In order to compare a formal notion of improvement with an intuitive one, we will attempt to develop a mechanism linking them in the same way as that in which the mechanism of elementary viewpoints links the formal notion of quality with the intuitive one. However, there must be an important difference between the mechanism which we have developed so far, on the one hand, and, on the other, the mechanism which we aim at in order to analyse improvements: a pair of distributions is not equivalent to any distribution, whereas a pair of improvements is an improvement. Points (1)–(5) below describe our mechanism.

(1) We can show that the transformation τ_1 completely defines the formal notion of improvement E^* implied by measure E. To this end, we show that two successive applications of τ_1 always yield the same change of E, thus yielding equal improvements.

(2) We can show that for every three distributions $\Delta_1 \cdot\!\geqslant \Delta_2 \cdot\!\geqslant \Delta_3$ there exist distributions $\Delta'_1, \Delta'_2, \Delta'_3$ and numbers m_1 and m_2 such that $\Delta_1, \Delta_2, \Delta_3$ are equivalent, from the viewpoint of E, to $\Delta_1, \Delta_2, \Delta_3$ with respect to $\Delta'_2 = \tau_1^{m_2}(\Delta'_3)$ and $\Delta'_1 = \tau_1^{m_1}(\Delta'_2)$. It can be shown that $\text{sign}(m_1 - m_2)$ is independent of the choice of $\Delta'_1, \Delta'_2, \Delta'_3$.

(3) Using (1) and (2), we may define the notion of the 'middle' of two distributions Δ_1 and Δ_3 as being a Δ_2 such that the sign function, as defined in (2), yields zero. This middle may be shown to fulfil the axioms of the bisymmetric structure as defined in Section 16.3.

(4) This, in turn, defines, according to measurement theory, an interval scale which in our case will coincide with the interval scale implied by measure E.

(5) In this way it will be shown that the notion of improvement connected with the transformation τ_1 defines unambiguously the notion of improvement connected with measure E. Hence, in order to find out whether or not some intuitive notion of improvement is compatible with the formal notion of improvement of measure E, we have to compare the intuitive notion with the transformation τ_1. If and only if our intuitive notion of improvement is met by τ_1 in the sense that successive applications of τ_1 always yield intuitively equivalent improvements, measure E may be used as interval scale.

In the same way as transformation τ_1 describes the improvement notion connected with measure E, the improvement notion of the measure F_1 of Jardine and van Rijsbergen (1971) may be described by, for example, the transformation τ_5: 'Replace two not-retrieved relevant documents by one retrieved relevant document'. τ_5 is obviously not equivalent to τ_1, which shows that E and F_1 are not equivalent as interval scales.

Analogously, the improvement notion of Vickery's measure F_2 may be shown to be described by the transformation τ_6: 'Replace a not-retrieved relevant document by two retrieved relevant documents.' τ_6 is obviously not equivalent either to τ_1 or τ_5. Hence, the measures E, F_1 and F_2 are pairwise not equivalent as interval scales.

16.5 Discussion of some problems

We now consider some of the problems mentioned in Section 16.2 insofar as they have not yet been considered. We start with problem (1) of Section 16.2.

We interpret question (1) as follows. Assume that the MZ-metric is defined (as a measure over some set $\mathscr{D}$ of objects to be measured) independently from the mechanism of additive conjoint measurement. Is it possible to analyse the MZ-metric into two partial measures — that is, is it possible to represent the MZ-metric as an additive conjoint measure? The answer depends on how the objects of $\mathscr{D}$ are defined and what the partial measures are that we have in mind.

Let $\mathscr{D}$ be the set of all distributions and let a, b, c, d be defined as in Section 16.2. Let $\mathscr{D}' \subset \mathscr{D}$ be the sub-set of $\mathscr{D}$ with $a \geqslant 1$ and $b+c \geqslant 1$. $\mathscr{D}'$ is represented as the Cartesian product $\mathscr{D}' = \mathscr{D}'_1 \times \mathscr{D}'_2$. We represent the contingency tables in the form

$$(a, b, c, d) = ((a, d), (b, c))$$

and define $\mathscr{D}'_1$ and $\mathscr{D}'_2$, respectively, as the sets $\mathscr{D}'_1 = \{(a, d) | a \geqslant 1\}$ and $\mathscr{D}_2 = \{(b, c) | b+c \geqslant 1\}$. We introduce on $\mathscr{D}'_1$ and $\mathscr{D}_2$ the following measures:

$$\varphi_1\colon \ \mathscr{D}'_1 \to \mathfrak{R} \quad \text{with } \varphi_1(a, d) = \log a$$

$$\varphi_2\colon \ \mathscr{D}'_2 \to \mathfrak{R} \quad \text{with } \varphi_2(b, c) = -\log(b+c)$$

The compound measure φ: $\mathscr{D}' \to \mathfrak{R}$ is defined by $\varphi = \varphi_1 + \varphi_2$. By use of the techniques of transformations, it can be shown that this measure is equivalent to the MZ-metric as an ordinal scale defined over the set of all distributions. From the above it follows that measure φ is an additive conjoint measure. The attempt to use recall and precision instead of φ_1 and φ_2 leads to difficulties, because a influences both recall and precision. In order still to use recall and precision we could proceed as follows. We represent the distribution in the form of a pair $((a, b), (r, d))$, where r is the recall. This representation is one-to-one for $r>0$ and $a \geqslant 1$.

Limiting ourselves to considering only distributions fulfilling this restriction, we then define φ_1 and φ_2 by

$$\varphi_1(a, b) = \frac{a}{a+b}$$

$$\varphi_2(r, d) = r$$

It is easily seen that φ_1 is precision and φ_2 is recall.

Question (2) may be interpreted in two ways. First, we may ask whether it is possible to have an additive conjoint measure based on recall and fallout. The answer is positive as far as technique is concerned. Consider the well-known measure $\frac{1}{2}(1+R-F)$ (Robertson, 1969). This measure is already presented in the form of a conjoint additive measure if we measure the recall on the distribution of all relevant documents and the fallout on the distribution of all non-relevant ones. The measured object is here the distribution represented as the product of two distributions: the distribution of all relevant documents and the distribution of all non-relevant documents. However, the question whether such a replacement is advantageous or even meaningful is not a technical one. To answer this question the underlying intuitive viewpoint must be specified and analysed.

The second interpretation for question (2) is: given the MZ-metric, is it possible to represent it as additive conjoint measure based on recall and fallout. The answer is negative, as by using the mechanism of transformations it may be shown that the MZ-metric, on the one hand, and the pair recall–fallout, on the other, possess transformational behaviours incompatible with their assumed additive relation.

Question (4), like question (2), can be interpreted in two ways. First, we may ask whether there exists some additive conjoint measure over the set of distributions, having three component measures. The answer is positive as far as technique is concerned. We represent the distribution one-to-one as triple $\Delta = ((G, N), r, f)$, where G is generality, N is the number of documents in the collection, r is the recall and f is the fallout. Then the precision can be represented as additive conjoint measure in the form

$$\psi\left(\log\frac{G}{1-G} + \log r - \log f\right)$$

where ψ is monotonic.

The second interpretation of the question is whether the MZ-metric may be represented as an additive conjoint measure with three component measures. The question remains open.

16.6 Conclusion

To conclude, we should like to remark that the mechanism of transformations and elementary viewpoints is a flexible and useful tool for analysing measures and their relation to viewpoints. It is a tool with a rather wide field of application, as already mentioned in the introduction. It is possible to describe with this technique such measures as recall, fallout, precision, normalised recall, recall–fallout graph and the paired measures recall and fallout as well as recall and precision.

References

BOLLMANN, P. and CHERNIAVSKY, V. S. (1979). 'Investigations on the *Evaluation Measures for Document-Retrieval-Systems by Transformations*, Technische Universität Berlin, Fachbereich Informatik, Bericht No. 77/26

BOLLMANN, P. and CHERNIAVSKY, V. S. (1979). 'Investigations on the recall–fallout graph by transformations' in *Proceedings of the 1979 Conference on Information Science and Systems*, Department of Electrical Engineering, The Johns Hopkins University, Baltimore

BOLLMANN, P. and CHERNIAVSKY, V. S. (1980). *Application of the Transformational Viewpoints' Mechanism to the MZ-Metric*, Technische Universität Berlin, Fachbereich Informatik, Bericht No. 80/10

CHERNIAVSKY, V. S. and LAKHUTY, D. G. (1970). 'Problem of evaluating retrieval systems I, *Naucho-Techniceskaya Informazia*, Ser. 2, 24–30 (in Russian): English translation, *Automatic Documentation and Mathematical Linguistics*, **4**, 9–26

HEINE, M. H. (1973). 'Distance between sets as an objective measure on retrieval effectiveness', *Information Storage and Retrieval*, **9**, 181–198

JARDINE, N. and VAN RIJSBERGEN, C. J. (1971). 'The use of hierarchic clustering in information retrieval', *Information Storage and Retrieval*, **7**, 217–240

KRANTZ, D. H., LUCE, R. D., SUPPES, P. and TVERSKY, A. (1971). *Foundations of Measurement*, Vol. 1, Academic Press, New York

LANCASTER, F. W. (1968). *Information Retrieval Systems*, Wiley, New York

PFANZAGL, J. (1973). *Theory of Measurement*, Würzburg

ROBERTSON, S. E. (1969). 'The parametric description of retrieval tests. Part 2: Overall measures', *Journal of Documentation*, **25**, 93–107

STEINHAUSEN, D. and LANGER, K. (1977). *Clusteranalyse*, Berlin

SWETS, J. A. (1969). 'Effectiveness of information retrieval methods', *American Documentation*, **20**, 72–89

VAN RIJSBERGEN, C. J. (1974). 'Foundation of evaluation', *Journal of Documentation*, **30**, 365–373

VAN RIJSBERGEN, C. J. (1979). *Information Retrieval* (2nd edn), Butterworths, London