An interval-like scale property for IR evaluation measures

Marco Ferrante
Dept. Mathematics
University of Padua, Italy
ferrante@math.unipd.it

Nicola Ferro
Dept. Information Engineering
University of Padua, Italy
ferro@dei.unipd.it

Silvia Pontarollo
Dept. Mathematics
University of Padua, Italy
spontaro@math.unipd.it

ABSTRACT

Evaluation measures play an important role in IR experimental evaluation and their properties determine the kind of statistical analyses we can conduct.

It has been previously shown that it is questionable that IR effectiveness measures are on an interval-scale and this implies that computing means and variances is not a permissible operation.

In this paper, we investigate whether it is possible to relax a bit the definition of interval scale, introducing the notion of interval-like scale, and to what extent IR effectiveness measures comply with this relaxed definition.

CCS CONCEPTS

•**Information systems → Retrieval effectiveness;**

KEYWORDS

evaluation measures; representational theory of measurement; interval scale

1 INTRODUCTION

Evaluation plays a central role in *Information Retrieval (IR)* and a lot of attention is devoted to improving our evaluation methodologies and practices. For example, since many years, there is a continued interest on how to properly apply statistical techniques to the analysis of IR experimental data, e.g., on the appropriate use of statistical testing [7, 13, 20, 23], on the normalization of measure values for cross-collection comparison [27], or on moving towards Bayesian inference [8, 21], just to name a few.

However, all these studies rely on some, often hidden and implicit, assumptions on what IR effectiveness measures are. In particular, *measurement scales* [15, 25] determine the operations that is admissible to perform with measure values and, as a consequence, the statistical analyses that can be applied. [25] identifies four major types of scales with increasing properties: (i) the *nominal scale* consists of discrete unordered values, i.e. categories; (ii) the *ordinal scale* introduces a natural order among the values; (iii) the *interval scale* preserves the equality of intervals or differences; and (iv) the *ratio scale* preserves the equality of ratios. Operations such as computing the mean or the variance are possible just on interval and ratio scales and they constitute the basis of many of the statistical techniques mentioned above. However, are we sure that IR effectiveness measures are on an interval scale? For example, [17] points out that the assumption of *Average Precision (AP)* being on an interval scale is somehow arbitrary and, as a consequence, also some of the descriptive statistics you compute about it.

Therefore, researchers started to study what IR effectiveness measures are, not only from an empirical perspective, e.g., [4, 5, 19], but also from a theoretical one, e.g., [1–3, 6, 10, 22, 26].

In this paper, we stem from the recent work of [11] and we move a step forward in understanding when and to what extent IR effectiveness measures are on an interval scale.

[11] investigated whether IR effectiveness measures are on an interval scale in the perspective of the *representational theory of measurement* [15], which is the measurement theory adopted in both physical and social sciences. According to this framework, the key point is to understand how real world objects, i.e., system runs in our case, are related to each other since measure properties are then derived from these relations. Moreover, it is important that these relations among real world objects are intuitive and sensible to "everybody" and that they can be commonly agreed on.

Therefore, [11] pointed out that the main issues in determining the scale of IR effectiveness measures are: (i) to understand how runs are empirically and intuitively ordered; (ii) to define what an interval of runs is; and, (iii) to determine how these intervals are ordered. Once you settled all these aspects, you can check whether an effectiveness measure comply with them or not and thus determine whether it is on an interval scale or not. In particular, [11] found that under a *strong top-heaviness* notion of ordering among runs, only *Rank-Biased Precision (RBP)* [16] with $p = \frac{1}{2}$ is on an interval scale while RBP for other values of p and other popular measures – namely AP, *Discounted Cumulated Gain (DCG)* [14], and *Expected Reciprocal Rank (ERR)* [9] – are not. Moreover, using a *weak top-heaviness* notion of ordering among runs, [11] found that all the previously mentioned IR effectiveness measures are not on an interval scale.

Strong top-heaviness provides us with a total ordering among runs and, as discussed above, there is at least one case of IR measure on an interval scale; however, the way in which strong top-heaviness orders runs may give raise to disagreement or corner cases. For example, strong top-heaviness ranks the run $(1, 0, 0, 0)$ with just one top relevant document before the run $(0, 1, 1, 1)$ with all relevant documents except for the first position; thus, there might be disagreement on whether this is an appropriate ordering for these runs. On the other hand, weak top-heaviness provides us with a much more intuitive partial ordering based on two basic operations – *swapping* two consecutive documents in a ranking and *replacing* a not relevant document with a relevant one [10]; however, none of the IR evaluation measures is on interval-scale using weak top-heaviness.

The problem with IR effectiveness measures emerging from [11] is two-fold: on the one side, both strong and weak top-heaviness create equi-spaced intervals of runs, as expected by the definition

8th International Workshop on Evaluating Information Access (EVIA 2017), co-located with NTCIR-13, 5 December 2017, Tokyo, Japan.

of interval scale, but IR effectiveness measures do not respect this equi-spacing; on the other side, both strong and weak top-heaviness do not account enough for the importance and the effect of the rank of a document in a run, since they both rely on the notion of *natural distance* in a poset (partially ordered set) [24] which flattens things too much, shrinking everything into a single number.

In this paper, we take a different approach to the ordering of intervals of runs, not based on single numbers, as the natural distance of [11] does, but using vectors instead. This new ordering is richer and more expressive than that induced by the natural distances in the strong and weak top-heaviness cases and allows us to introduce the notion of *interval-like scale*, i.e., something richer than an ordinal scale but a bit less powerful than an interval scale, since runs are ordered, intervals of runs are ordered too but intervals may not be equi-spaced. In particular, we find that, under reasonable assumptions, DCG and RBP are on a interval-like scale while AP and ERR are not.

The paper is organized as follows: Section 2 recaps some basic concepts about the representational theory of measurement and posets; Section 3 deals with interval-like scales; finally, Section 4 wraps up the discussion and outlooks some future work.

2 BACKGROUND

2.1 Representational Theory of Measurement

A **relational structure** [15, 18] is an ordered pair $\mathbf{X} = \langle X, R_X \rangle$ of a domain set X and a set of relations R_X on X, where the relations in R_X may have different arities, i.e. they can be unary, binary, ternary relations and so on. Given two relational structures $\mathbf{X}$ and $\mathbf{Y}$, a *homomorphism* $\mathbf{M} : \mathbf{X} \to \mathbf{Y}$ from $\mathbf{X}$ to $\mathbf{Y}$ is a mapping $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ where: (i) M is a function that maps X into $\mathrm{M}(X) \subseteq Y$, i.e. for each element of the domain set there exists one corresponding image element; (ii) M_R is a function that maps R_X into $\mathrm{M}_R(R_X) \subseteq R_Y$ such that $\forall r \in R_X$, r and $\mathrm{M}_R(r)$ have the same arity, i.e. for each relation on the domain set there exists one (and it is usually, and often implicitly, assumed: and only one) corresponding image relation; (iii) $\forall r \in R_X, \forall x_i \in X$, if $r(x_1, \ldots, x_n)$ then $\mathrm{M}_R(r)\Big(\mathrm{M}(x_1), \ldots, \mathrm{M}(x_n)\Big)$, i.e. if a relation holds for some elements of the domain set then the image relation must hold for the image elements.

A relational structure $\mathbf{E}$ is called *empirical* if its domain set E spans over the entities under consideration in the real world, i.e. the system runs in our case; a relational structure $\mathbf{S}$ is called *symbolic* if its domain set S spans over a given set of numbers. A **measurement (scale)** is the homomorphism $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ from the real world to the symbolic world and a **measure** is the number assigned to an entity by this mapping.

2.2 Measurement Scales

[11] relied on the notion of *difference structure* [15, 18] to introduce a definition of interval among system runs in such a way that it ensures the existence of an interval scale.

Given E, a weakly ordered empirical structure is a pair $(E, \preceq)$ where, for every $a, b, c \in E$,

- $a \preceq b$ or $b \preceq a$;
- $a \preceq b$ and $b \preceq c \implies a \preceq c$.

Given $(E, \preceq)$, we have to define a **difference** Δ_{ab} between two elements $a, b \in E$, which is a kind of signed distance we exploit to compare intervals. Then, we have to define a weak order $\preceq_d$ between these Δ_{ab} differences. We can proceed as follows: if two elements $a, b \in E$ are such that $a \sim b$, i.e. $a \preceq b$ and $b \preceq a$, then the interval $[a, b]$ is null and, consequently, we set $\Delta_{ab} \sim_d \Delta_{ba}$; if $a \succ b$ we agree upon choosing $\Delta_{aa} \prec_d \Delta_{ab}$ which, in turn implies that $\Delta_{aa} \succ_d \Delta_{ba}$.

Definition 1. Let E be a finite (not empty) set of objects. Let $\preceq_d$ be a binary relation on $E \times E$ that satisfies, for each $a, b, c, d, a', b', c' \in E$, the following axioms:

i. $\preceq_d$ is *weak order*;
ii. if $\Delta_{ab} \preceq_d \Delta_{cd}$, then $\Delta_{dc} \preceq_d \Delta_{ba}$;
iii. if $\Delta_{ab} \preceq_d \Delta_{a'b'}$ and $\Delta_{bc} \preceq_d \Delta_{b'c'}$ then $\Delta_{ac} \preceq_d \Delta_{a'c'}$;
iv. *Solvability Condition*: if $\Delta_{aa} \preceq_d \Delta_{cd} \preceq_d \Delta_{ab}$, then there exists $d', d'' \in E$ such that $\Delta_{ad'} \sim_d \Delta_{cd} \sim_d \Delta_{d''b}$.

Then $(E, \preceq_d)$ is a **difference structure**.

Particular attention has to be paid to the *Solvability Condition* which ensures the existence of an equally spaced gradation between the elements of E, indispensable to construct an interval scale measurement.

The *representation theorem* for difference structures states:

Theorem 1. *Let E be a finite (not empty) set of objects and let $(E, \preceq_d)$ be a difference structure. Then there exist a measurement scale $\mathrm{M} : E \to \mathbb{R}$ such that for every $a, b, c, d \in E$*

$$\Delta_{ab} \preceq_d \Delta_{cd} \Leftrightarrow \mathrm{M}(a) - \mathrm{M}(b) \leq \mathrm{M}(c) - \mathrm{M}(d) \ .$$

This theorem ensures us that, if there is a difference structure on the empirical set E, then there exists an interval scale M.

As anticipated in Section 1, we will introduce the notion of *interval-like scale* which corresponds to removing the solvability condition from the definition of difference structure and obtaining a new partial ordering of the intervals of runs.

2.3 Posets

A partially ordered set P, **poset** for short, is a set with a partial order $\preceq$ defined on it [24]. A **partial order** $\preceq$ is a binary relation over P which is reflexive, antisymmetric and transitive. Given $s, t \in P$, we say that s and t are *comparable* if $s \preceq t$ or $t \preceq s$, otherwise they are *incomparable*.

A closed **interval** is a subset of P defined as $[s, t] := \{u \in P : s \preceq u \preceq t\}$, where $s, t \in P$ and $s \preceq t$. Moreover we say that t **covers** s if $s \preceq t$ and $[s, t] = \{s, t\}$, that is there does not exist $u \in P$ such that $s \prec u \prec t$.

We can represent a finite poset P by using the **Hasse diagram** which is a graph where vertices are the elements of P, edges represent the *covers* relations, and if $s \prec t$ then s is below t in the diagram.

A subset C of a poset P is a **chain** if any two elements of C are comparable: a chain is a totally ordered subset of a poset. If C is a finite chain, the **length** of C, $\ell(C)$, is defined by $\ell(C) = |C| - 1$. A **maximal chain** of P is a chain that is not a proper subset of any other chain of P.

If every maximal chain of P has the same length n, we say that P is **graded of rank n**; in particular there exists a unique function

$\rho : P \to \{0, 1, \ldots, n\}$, called the **rank function**, such that $\rho(s) = 0$, if s is a minimal element of P, and $\rho(t) = \rho(s) + 1$, if t covers s.

Finally, since any interval on a graded poset is graded, the **length of an interval** $[s, t]$ is given by $\ell(s, t) := \ell([s, t]) = \rho(t) - \rho(s)$, also called the **natural distance**.

3 INTERVAL-LIKE SCALES

3.1 Preliminary Definitions

Given N, the length of the run, we define the **set of retrieved documents** as $D(N) = \{(d_1, \ldots, d_N) : d_i \in D, d_i \neq d_j \text{ for any } i \neq j\}$, i.e. the ranked list of retrieved documents without duplicates, and the **universe set of retrieved documents** as $\mathcal{D} := \bigcup_{N=1}^{|D|} D(N)$. A **run** r_t, retrieving a ranked list of documents $D(N)$ in response to a topic $t \in T$, is a function from T into $\mathcal{D}$

$$t \mapsto r_t = (d_1, \ldots, d_N)$$

We denote by $r_t[j]$ the j-th element of the vector r_t, i.e. $r_t[j] = d_j$.

We define the **universe set of judged documents** as $\mathcal{R} := \bigcup_{N=1}^{|D|} REL^N$, where REL^N is the set of the ranked lists of judged retrieved documents with length fixed to N. Since in our case $REL = \{0, 1\}$, $REL^N = \{0, 1\}^N$ refers to the space of all $N-$length vectors consisting of 0 and 1. As for the set-based case, we denote by RB_t the **recall base**, i.e. the total number of relevant documents for a topic.

We call **judged run** the function $\hat{r}_t$ from $T \times \mathcal{D}$ into $\mathcal{R}$, which assigns a relevance degree to each retrieved document in the ranked list

$$(t, r_t) \mapsto \hat{r}_t = \big(GT(t, d_1), \ldots, GT(t, d_N)\big)$$

We denote by $\hat{r}_t[j]$ the j-th element of the vector $\hat{r}_t$, i.e. $\hat{r}_t[j] = GT(t, d_j)$.

As for the set-based case, we can simplify the notation omitting the dependence on topics, $\hat{r} := \big(\hat{r}[1], \ldots, \hat{r}[N]\big)$, RB, and so on.

3.2 Ordering between Intervals

Let us start recalling the ordering between runs adopted in this paper and based on the following two *monotonicity-like* properties proposed by [10]:

- **Replacement** A measure of retrieval effectiveness should not decrease when replacing a document with another one in the same rank position with higher degree of relevance.
- **Swap** If we swap a less relevant document with a more relevant one in a lower rank position, the measure should not decrease.

These two properties lead to the following partial ordering among system runs

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \sum_{j=1}^{k} \hat{r}[j] \leq \sum_{j=1}^{k} \hat{s}[j] \;\; \forall k \in \{1, \ldots, N\} \,. \tag{1}$$

This ordering considers a run bigger than another one when, for each rank position, it has more relevant documents than the other one up to that rank.

This is the same ordering of runs used by [11] in the weak top-heaviness case but, differently from [11], we now introduce a different notion of length of an interval, not based on the natural distance which, as discussed in Section 1, has the drawback of flattening everything into a single number.

To define the length of an interval we adopt the following strategy: given $\hat{r}, \hat{s} \in REL^N$ with $\hat{r} \preceq \hat{s}$, we count how many replacements in the last position and how many forward single-step swaps at each depth are necessary to go from $\hat{r}$ to $\hat{s}$ following a maximal chain in REL^N. In order to do this, it is useful to define the cumulative sums of a vector $v = (v[1], \ldots, v[N])$, denoted using the capital letter as $V = (V[1], \ldots, V[N])$, where $V[j] = \sum_{i=1}^{j} v[i]$.

Let us start with a simple example.

Example. Consider the two judged runs in REL^4

$$\hat{r} = (0, 1, 1, 0) \,,$$
$$\hat{0} = (0, 0, 0, 0) \,.$$

Since $\hat{0} \prec \hat{r}$, in order to construct a chain from $\hat{0}$ to $\hat{r}$ with the two basic operators (replacement in last position and single-step forward swap) we get

$$\hat{0} = (0, 0, 0, 0) \,,$$
$$\hat{0}_1 = (0, 0, 0, 1) \,,$$
$$\hat{0}_2 = (0, 0, 1, 0) \,,$$
$$\hat{0}_4 = (0, 1, 0, 0) \,,$$
$$\hat{0}_5 = (0, 1, 0, 1) \,,$$
$$\hat{0}_6 = (0, 1, 1, 0) = \hat{r} \,.$$

We have made two replacement in the fourth position, one swap in the second position and two in the third one. Recall that with *swap at depth i* we mean that a forward swap from position $i - 1$ to position i was done. We can *count* how many of these basic operations in each position are needed to go from $\hat{0}$ to $\hat{r}$ just taking the cumulative sums of $\hat{r}$. Indeed we get

$$\hat{R} = (0, 1, 2, 2) \,,$$

and each entry $k < D$ of $\hat{R}$, $\hat{R}[k]$, counts the number of swaps made in position k, while $\hat{R}[N]$ counts the number of replacement, i.e. the total mass of $\hat{r}$, to go from $\hat{0}$ to $\hat{r}$.

More generally, given two vectors $\hat{r}, \hat{s} \in REL^N$, with $\hat{r} \prec \hat{s}$, in order to collect the number of basic operations made at each position to go from $\hat{r}$ to $\hat{s}$, we can compute this vector of length N first between $\hat{0}$ and $\hat{r}$ and between $\hat{0}$ and $\hat{s}$, namely $\hat{R}$ and $\hat{S}$, and then subtract the two vectors. Precisely $\hat{S} - \hat{R}$ leads to a new vector of length N, where each entry k equals the number of swaps or replacements (if $k = N$) needed to go from $\hat{r}$ to $\hat{s}$.

Example. In order to better understand this mechanism, let us consider a second example. Consider the two judged runs in REL^4

$$\hat{r} = (0, 1, 0, 0) \,,$$
$$\hat{s} = (1, 0, 1, 0) \,.$$

In order to construct a chain from $\hat{r}$ to $\hat{s}$ with the two basic operators (replacement in last position and single-step forward swap) we get

$$\hat{r} = (0, 1, 0, 0) \,,$$
$$\hat{v} = (1, 0, 0, 0) \,,$$
$$\hat{w} = (1, 0, 0, 1) \,,$$
$$\hat{s} = (1, 0, 1, 0) \,.$$

We have made a swap in the first and third position and a replacement in the fourth position, that we can collect in a vector as

$$t = (1, 0, 1, 1)\ . \tag{2}$$

On the other hand it is easy to compute $\hat{R} = (0, 1, 1, 1)$ and $\hat{S} = (1, 1, 2, 2)$. Therefore

$$\hat{S} - \hat{R} = (1, 0, 1, 1) = t\ ,$$

as we wanted to show.

Let us consider a second, more complicated, example.

Example. Consider the two judged runs

$$\hat{r} = (1, 0, 0, 0, 0, 1, 1, 0, 1, 0)\ ,$$
$$\hat{s} = (1, 1, 0, 0, 1, 0, 1, 0, 0, 1)\ .$$

Clearly $\hat{r} \preceq \hat{s}$. Moreover

$$\hat{S} = (1, 2, 2, 2, 3, 3, 4, 4, 4, 5)\ ,$$
$$\hat{R} = (1, 1, 1, 1, 1, 2, 3, 3, 4, 4).$$

Thus

$$\hat{S} - \hat{R} = (0, 1, 1, 1, 2, 1, 1, 1, 0, 1).$$

Let $t = \hat{S} - \hat{R}$. For any $i < 10$, $t[i]$ tells us how many swaps one needs to do at depth i to make the smallest run coincide with the biggest one. Moreover, if the total number of relevant relevance-degrees is not equal for both, as in this example, the last entry of t, $t[N]$, is exactly the number or replacements on $\hat{r}$ one needs to make, and coincide with $\sum_i \hat{s}(i) - \sum_i \hat{r}(i)$.

Given an interval $[\hat{r}, \hat{s}]$, if we take the cumulative sums of $t = \hat{S} - \hat{R}$ we obtain the vector T of the cumulative sums of t that counts, for every $i \leq N$, the total number of swaps (or replacements, if $i = N$) made from depth 1 to i between the endpoints of the given interval. The vector T can be seen as a new and generalized definition of the **length** of the interval $[\hat{r}, \hat{s}]$, which replaces the *natural distance* used by [11].

According to this new distance, we say that the interval $[\hat{r_1}, \hat{s_1}]$ is smaller than or equal to the interval $[\hat{r_2}, \hat{s_2}]$ if, for the vectors T_1 and T_2 of their cumulative sums, it holds that $T_1[i] \leq T_2[i]$ for any $i \leq n$. It is worth noticing that, if we take as definition of length any convex linear combination of the values $(T[i], \dots, T[n])$, the intervals comparable for the previous ordering remain comparable. Other intervals become comparable for any fixed linear combination, but it is not possible to say in advance they are ordered in the same way by any two of these combinations.

We are now able to define a difference in this setting:

Definition 2. Given $\hat{r}, \hat{s} \in REL^N$, with $\hat{r} \preceq \hat{s}$, the **difference** $\vec{\Delta}_{\hat{s}\hat{r}}$ is a vector of length N such that

$$\vec{\Delta}_{\hat{s}\hat{r}}[i] \coloneqq \sum_{j=1}^{i} (i - j + 1)\big(\hat{s}[j] - \hat{r}[j]\big)\ ,$$

for all $i \in \{1, \dots, N\}$.

It can be easily proved that $\vec{\Delta}_{\hat{s}\hat{r}}$ is exactly the vector T defined above. Indeed, by construction, given $\hat{r}, \hat{s} \in REL^N$ with $\hat{r} \preceq \hat{s}$, $t[j] = \sum_{n=1}^{j} (\hat{s}[n] - \hat{r}[n])$. Therefore $T[i] = \sum_{j=1}^{i} t[j] = \sum_{j=1}^{i} \sum_{n=1}^{j} (\hat{s}[n] - \hat{r}[n]) = \sum_{j=1}^{i} (i - j + 1)\big(\hat{s}[j] - \hat{r}[j]\big)$.

Moreover, when computing the difference vector $\vec{\Delta}..$ between two comparable runs $\hat{r}, \hat{s}$, in this work we write $\vec{\Delta}_{\hat{s}\hat{r}}$ whenever $\hat{r} \preceq \hat{s}$: if we instead consider $\vec{\Delta}_{\hat{r}\hat{s}}$, then we are counting the backward swaps from $\hat{s}$ to $\hat{r}$ and $\vec{\Delta}_{\hat{r}\hat{s}}[i] \leq 0$ for all $i \in \{1, \dots, N\}$.

Since here $\vec{\Delta}..$ is no more a scalar but a vector, we have to define the partial order among intervals of runs $\preceq_d$ as follow:

Definition 3. Given $[\hat{r}, \hat{s}], [\hat{u}, \hat{v}] \subseteq REL^N$,

$$\vec{\Delta}_{\hat{v}\hat{u}} \preceq_d \vec{\Delta}_{\hat{s}\hat{r}}$$

if and only if

$$\vec{\Delta}_{\hat{v}\hat{u}}[i] \leq \vec{\Delta}_{\hat{s}\hat{r}}[i], \quad \forall i \in \{1, \dots, N\}.$$

Example. With respect to the previous example, where $t = \hat{S} - \hat{R} = (0, 1, 1, 1, 2, 1, 1, 1, 0, 1)$, the vector $\vec{\Delta}_{\hat{s}\hat{r}}$ is given by

$$\vec{\Delta}_{\hat{s}\hat{r}} = T = (0, 1, 2, 3, 5, 6, 7, 8, 8, 9).$$

Let now $\hat{u}, \hat{v} \in \{0, 1\}^{10}$ be as follows

$$\hat{u} = (1, 0, 0, 1, 0, 1, 1, 1, 0, 0)\ ,$$
$$\hat{v} = (1, 0, 1, 1, 1, 0, 1, 0, 0, 0)\ .$$

Clearly $\hat{u} \preceq \hat{v}$ and

$$\vec{\Delta}_{\hat{v}\hat{u}} = (0, 0, 1, 2, 4, 5, 6, 6, 6, 6)\ .$$

Thus we can conclude that the difference between $\hat{s}$ and $\hat{r}$ is greater than the difference between $\hat{v}$ and $\hat{u}$.

Note that the last entry of $\vec{\Delta}..$ always equals the natural distance as defined in Section 2.3 and used by [11]. Indeed, given two comparable runs $\hat{r}, \hat{s} \in REL^N$, with $\hat{r} \preceq \hat{s}$, $\vec{\Delta}_{\hat{s}\hat{r}}[N]$ counts the total number of forward swaps of length one and/or replacements done from $\hat{r}$ to match $\hat{s}$. Since swaps of length one and replacements in the last positions are elementary operations as observed above, then $\vec{\Delta}_{\hat{s}\hat{r}}[N]$ is just counting the length of every maximal chain in $[\hat{r}, \hat{s}]$, i.e., exactly the natural distance.

This definition of *difference vector* solves some of the problems encountered with the difference defined using the natural distance, as the following example shows.

Example. Let $\hat{r}, \hat{s}, \hat{u}, \hat{v}$ be defined as follows:

$$\hat{r} = (0, \mathbf{1}, 0, 0, 0, 0, 0, 0, 0, 0)\ ,$$
$$\hat{s} = (\mathbf{1}, 0, 0, 0, 0, 0, 0, 0, 0, 0)\ ,$$
$$\hat{u} = (0, 0, 0, 0, 0, 0, 0, 0, 0, \mathbf{1})\ ,$$
$$\hat{v} = (0, 0, 0, 0, 0, 0, 0, 0, \mathbf{1}, 0)\ ,$$

where $\hat{r} \preceq \hat{s}$ and $\hat{u} \preceq \hat{v}$.

As already discussed, the natural distance induces a difference between runs that does not keep track or the rank. In this case, the natural distance would that both the pairs $\hat{r}, \hat{s}$, and $\hat{u}, \hat{v}$, have both difference equal to 1, even if these two pair differs a lot in terms of where differences actually happen in the ranking.

Instead, $\vec{\Delta}..$ shows a bigger difference between $\hat{r}$ and $\hat{s}$ compared to the other two runs, because their differences happen in higher and more important rank positions:

$$\vec{\Delta}_{\hat{s}\hat{r}} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$$
$$\vec{\Delta}_{\hat{v}\hat{u}} = (0, 0, 0, 0, 0, 0, 0, 0, 1, 1)\ ,$$

and $\vec{\Delta}_{\hat{s}\hat{r}}[i] \geq \vec{\Delta}_{\hat{v}\hat{u}}[i]$ for every $i \in \{1, \dots, 10\}$.

Therefore, this new and more expressive difference matches better with the intuition that the higher the rank position at which it happens, the more important the same difference between two runs.

The vector $\vec{\Delta}..$ is thus useful to compare, when possible, intervals on REL^N, paying the necessary attention on the ranking. As a consequence, a measure that satisfy these relations among intervals, although not interval scale, could be viewed as something more powerful than a measure on ordinal scale. Indeed, when the above differences between intervals are comparable, one direction of *iff* on Theorem 1 is still satisfied.

Therefore we can say that a measure M of retrieval effectiveness is **interval-like** if, given a distance (potentially vector) $\Delta..$, an ordering $\preceq_d$ between distances, and given $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in REL^N$, the following relation holds:

$$\Delta_{\hat{s}\hat{r}} \preceq_d \Delta_{\hat{v}\hat{u}} \Rightarrow \mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) \leq \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u}).$$

The next section is discusses whether some well-known IR measures are *interval-like* with respect to the difference introduced in Definition 2.

3.3 Interval-like Scale Measures

We tested some measures of retrieval effectiveness – namely AP, RBP_p, ERR, DCG – on intervals with comparable differences according to the above definition.

ERR shows the strongest discordance with our definition of difference, since often it does not respect the relations between intervals induced by $\vec{\Delta}..$, as the next example shows.

Example. Let us consider the following four runs $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in \{0,1\}^{10}$:

$$\hat{r} = (0,0,0,0,0,0,1,1,1,0)\,,$$
$$\hat{s} = (0,0,0,0,0,1,0,1,1,0)\,,$$
$$\hat{u} = (1,1,0,1,0,1,1,0,1,1)\,,$$
$$\hat{v} = (1,1,1,0,0,1,1,0,1,1)\,.$$

Clearly $\hat{r} \preceq \hat{s} \preceq \hat{u} \preceq \hat{v}$. It seems fair to think that $\hat{r}$ and $\hat{s}$ give rise to a smaller interval compared to $[\hat{u}, \hat{v}]$ – note that the endpoints of both intervals differ by a swap of length one, but made in different positions. Moreover it is easy to prove that $\vec{\Delta}_{\hat{s}\hat{r}}[i] \leq \vec{\Delta}_{\hat{v}\hat{u}}[i]\ \forall i$. But while the measures RBP_p, AP and DCG agree with the previous statement, ERR does not, since $ERR(\hat{s}) - ERR(\hat{r}) > ERR(\hat{v}) - ERR(\hat{u})$.

Another measure that does not always respect the relations between distances is AP.

Example. Let us consider the following runs $\hat{r}, \hat{s}, \hat{u} \in \{0,1\}^{10}$:

$$\hat{r} = (0,0,0,0,0,0,0,0,0,0)\,,$$
$$\hat{s} = (0,1,0,0,1,0,0,0,0,1)\,,$$
$$\hat{u} = (0,1,0,0,1,1,1,0,0,1)\,.$$

Clearly $\hat{r} \preceq \hat{s}$ and $\hat{s} \preceq \hat{u}$. The readers can agree to consider the interval $[\hat{r}, \hat{s}]$ strictly bigger than $[\hat{s}, \hat{u}]$, since from $\hat{u}$ to $\hat{s}$ we have lost only two relevant documents, while from $\hat{s}$ to $\hat{r}$ the information lost seems to be higher. Moreover $\vec{\Delta}_{\hat{s}\hat{r}}[i] \geq \vec{\Delta}_{\hat{u}\hat{s}}[i]\ \forall i$, with strict inequality for some i. However while the measures RBP_p, ERR and DCG agree with this relation between the two intervals, AP does not, since $AP(\hat{s}) - AP(\hat{r}) < AP(\hat{u}) - AP(\hat{s})$.

Instead, RBP_p and DCG show a greater agreement with the inequalities between intervals induced by $\vec{\Delta}..$, even if sometimes they do not respect these relations: this happens when the endpoints of an interval do not have an equal number of relevant documents.

Example. Let us consider $\hat{r}, \hat{s}, \hat{u} \in \{0,1\}^{10}$:

$$\hat{r} = (0,0,1,0,1,1,0,0,1,0)\,,$$
$$\hat{s} = (0,1,0,1,0,1,1,1,1,0)\,,$$
$$\hat{u} = (1,1,0,1,1,1,0,1,0,0)\,.$$

Clearly $\hat{r} \preceq \hat{s} \preceq \hat{u}$ and one can prove that

$$\vec{\Delta}_{\hat{s}\hat{r}} = (0,1,1,2,2,2,3,5,7,9)\,,$$
$$\vec{\Delta}_{\hat{u}\hat{s}} = (1,2,3,4,6,8,9,10,10,10)\,,$$

that is $\vec{\Delta}_{\hat{s}\hat{r}}[i] \leq \vec{\Delta}_{\hat{u}\hat{s}}[i]\ \forall i$, with strict inequality for some i. While $\hat{u}$ and $\hat{s}$ has the same number of relevant documents, $\hat{r}$ has two relevant documents less than $\hat{s}$. In particular $DCG(\hat{s}) - DCG(\hat{r}) > DCG(\hat{u}) - DCG(\hat{s})$ and, for $p > 0.85$, $RBP_p(\hat{s}) - RBP_p(\hat{r}) > RBP_p(\hat{u}) - RBP_p(\hat{s})$, against the inequality given by the difference vectors.

Therefore, we can say that RBP_p and DCG are *interval-like* with respect to the difference introduced in Definition 2 and considering only intervals where the endpoints have an equal number of relevant documents. While AP and ERR are not even *interval-like* since the relations between intervals often fail to be complied with.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we conducted a formal study to propose a new and more expressive way of providing an empirical ordering of intervals of runs in order to determine how close IR effectiveness measure are to be on an interval scale. Indeed, previous work [10, 11] has shown that they are on an ordinal scale, under some conditions, but not on an interval scale. We have introduced the notion of interval-like scale, a kind of interval scale which admits intervals to not be equi-spaced, and we have shown that both DCG and RBP are on this scale, under reasonable conditions, while AP and ERR are not.

Future work will concern an empirical investigation of the different theoretical properties of evaluation measures we have found in order to determine the impact and severity of not complying with them when you compute descriptive statistics, like mean and variance, and when you conduct statistical significance tests.

REFERENCES

[1] E. Amigó, J. Gonzalo, and M. F. Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks. In *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai (Eds.). ACM Press, New York, USA, 643–652.

[2] P. Bollman. 1984. Two Axioms for Evaluation Measures in Information Retrieval. In *Proc. of the Third Joint BCS and ACM Symposium on Research and Development in Information Retrieval*, C. J. van Rijsbergen (Ed.). Cambridge University Press, UK, 233–245.

[3] P. Bollmann and V. S. Cherniavsky. 1980. Measurement-theoretical investigation of the MZ-metric. In *Proc. 3rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1980)*, C. J. van Rijsbergen (Ed.). ACM Press, New York, USA, 256–267.

[4] C. Buckley and E. M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, E. Yannakoudakis, N. J. Belkin, M.-K. Leong, and P. Ingwersen (Eds.). ACM Press, New York, USA, 33–40.

[5] C. Buckley and E. M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza (Eds.). ACM Press, New York, USA, 25–32.

[6] L. Busin and S. Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proc. 4th International Conference on the Theory of Information Retrieval (ICTIR 2013)*, O. Kurland, D. Metzler, C. Lioma, B. Larsen, and P. Ingwersen (Eds.). ACM Press, New York, USA, 22–29.

[7] B. A. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 4:1–4:34.

[8] B. A. Carterette. 2015. Bayesian Inference for Information Retrieval Evaluation. In *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang (Eds.). ACM Press, New York, USA, 31–40.

[9] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin (Eds.). ACM Press, New York, USA, 621–630.

[10] M. Ferrante, N. Ferro, and M. Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang (Eds.). ACM Press, New York, USA, 21–30.

[11] M. Ferrante, N. Ferro, and S. Pontarollo. 2017. Are IR Evaluation Measures on an Interval Scale?. In *Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017)*, J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz (Eds.). ACM Press, New York, USA, 67–74.

[12] S. Foldes. 2013. On distances and metrics in discrete ordered sets. *arXiv.org, Combinatorics (math.CO)* arXiv:1307.0244 (June 2013).

[13] D. A. Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, R. Korfhage, E. Rasmussen, and P. Willett (Eds.). ACM Press, New York, USA, 329–338.

[14] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (October 2002), 422–446.

[15] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement. Additive and Polynomial Representations.* Vol. 1. Academic Press, New York, USA.

[16] A. Moffat and J. Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2:1–2:27.

[17] S. Robertson. 2006. On GMAP: and Other Transformations. In *Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006)*, P. S. Yu, V. Tsotras, E. A. Fox, and C.-B. Liu (Eds.). ACM Press, New York, USA, 78–83.

[18] G. B. Rossi. 2014. *Measurement and Probability. A Probabilistic Theory of Measurement with Applications.* Springer-Verlag, New York, USA.

[19] T. Sakai. 2006. Evaluating Evaluation Metrics based on the Bootstrap. In *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, E. N. Efthimiadis, S. Dumais, D. Hawking, and K. Järvelin (Eds.). ACM Press, New York, USA, 525–532.

[20] T. Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (June 2014), 3–12.

[21] T. Sakai. 2017. The Probability that Your Hypothesis Is Correct, Credible Intervals, and Effect Sizes for IR Evaluation. In *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White (Eds.). ACM Press, New York, USA, 25–34.

[22] F. Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. In *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang (Eds.). ACM Press, New York, USA, 11–20.

[23] M. D. Smucker, J. Allan, and B. A. Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proc. 16th International Conference on Information and Knowledge Management (CIKM 2007)*, M. J. Silva, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. and Falcão (Eds.). ACM Press, New York, USA, 623–632.

[24] R. P. Stanley. 2012. *Enumerative Combinatorics – Volume 1* (2nd ed.). Cambridge Studies in Advanced Mathematics, Vol. 49. Cambridge University Press, Cambridge, UK.

[25] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science, New Series* 103, 2684 (June 1946), 677–680.

[26] C. J. van Rijsbergen. 1974. Foundations of Evaluation. *Journal of Documentation* 30, 4 (1974), 365–373.

[27] W. Webber, A. Moffat, and J. Zobel. 2008. Score Standardization for Inter-Collection Comparison of Retrieval Systems. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani (Eds.). ACM Press, New York, USA, 51–58.