Are IR Evaluation Measures on an Interval Scale?

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

In this paper, we formally investigate whether, or not, IR evaluation measures are on an interval scale, which is needed to safely compute the basic statistics, such as mean and variance, we daily use to compare IR systems. We face this issue in the framework of the representational theory of measurement and we rely on the notion of difference structure, i.e. a total equi-spaced ordering on the system runs.

We found that the most popular set-based measures, i.e. precision, recall, and F-measure are interval-based. In the case of rank-based measures, using a strongly top-heavy ordering, we found that only RBP with $p = \frac{1}{2}$ is on an interval scale while RBP for other p values, AP, DCG, and ERR are not. Moreover, using a weakly top-heavy ordering, we found that none of RBP, AP, DCG, and ERR is on an interval scale.

CCS CONCEPTS

• **Information systems → Retrieval effectiveness**;

KEYWORDS

evaluation measures; representational theory of measurement; interval scale

1 INTRODUCTION

Information Retrieval (IR) is deeply rooted in experimentation but there is a growing need for stronger theoretical foundations [9]. Even if experimental evaluation is a main driver of progress and IR measures are a core part of it, our theoretical understanding of what IR measures are is still quite limited, despite the several studies both in the past [3, 4, 21] and more recently [2, 5, 7, 18].

When measuring something, the notion of *measurement scale* plays a central role [12, 20], since it determines the operations that can be performed and, as a consequence, the statistical analyses that can be applied. Stevens [20] 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.

ICTIR '17, October 01-04, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-4490-6/17/10…$15.00
https://doi.org/10.1145/3121050.3121058

Many of the operations we perform daily in IR, such as computing averages and variances, are possible only from interval scales onwards but, due to our limited knowledge of IR measures, we do not actually know which scales they rely on. Robertson [16] points out that the assumption of *Average Precision (AP)* being on an interval scale is somehow arbitrary while Ferrante et al. [7] shows that, only under some strong restrictions on the system runs being compared, we can ensure that IR measures use at least an ordinal scale.

We investigate whether IR measures are on an interval scale or not. We rely on the *representational theory of measurement* [12], 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 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.

In our context, this means that being on an interval scale is not just a numeric property of an IR measure but we need first to understand how system runs are ordered and what *intervals* of system runs are. Then, once we come to commonly agreed notions of order and interval among system runs, we can verify whether an IR measure complies with these notions and determine whether it is on an interval scale or not.

Therefore, we introduce a notion of interval among system runs by relying on *posets* and *lattices* [19] and we exploit *Hasse diagrams* to provide a graphical representation of such intervals. Then, we define a *difference* which quantifies the "length" of such intervals.

In the case of set-based measures and system runs of fixed length, we show that our notions of interval and difference induce a *difference structure* [12, 17] on the set of system runs and this guarantees the existence of an interval scale measure M while a uniqueness theorem ensures that any other interval scale measure is just a positive linear transformation of such M. We then show how to construct such interval scale measure M and we prove that Precision, Recall, and F-measure are all on an interval scale by finding a positive linear transformation with such measure M.

In the case of rank-based measures and system runs of fixed length, we explore two different notions of interval and difference based on the *top-heaviness* property, i.e. the preference towards highly ranked relevant documents. Using a *strong top-heaviness* notion, we find how it induces a difference structure and we prove that only *Rank-Biased Precision (RBP)* [15] 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)* [10], and *Expected Reciprocal Rank (ERR)* [6] – are not. Using a *weak top-heaviness* notion, we find that it induces another difference

structure and we prove that the previously mentioned IR measures are not on an interval scale.

The paper is organized as follows: Section 2 introduces some basic concepts about the representational theory of measurement and how to determine if a measure is on an interval scale; Section 3 recalls some definition and properties of posets and Hasse diagrams; Section 4 analyses set-based IR measures while Section 5 deals with rank-based IR measures; finally, Section 6 wraps up the discussion and outlooks some future work.

2 MEASUREMENT THEORY

2.1 Representational Theory of Measurement

The *representational theory of measurement* [12] sees measurement as the process of assigning numbers to the entities in the real world according to some property under examination. Therefore, the relations among the entities in the real world determine the relations among the numbers we assign.

More precisely, a **relational structure** [12, 17] 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

As discussed in Section 1, there are four major types of measurement scales [20] which can be ordered by their increasing properties and allows for different computations: *nominal scales* allow us to compute the number of cases and the mode; in addition, *ordinal scales* allow us to compute median and percentiles; *interval scales* add the possibility to compute mean, variance, product-moment correlation and rank correlation; finally, *ratio scales* add the capability to compute the coefficient of variation. Over the years, there has been debate [22] on whether these rules are too strict or not but they are applied widely.

If we already know that on an empirical structure there is an interval scale M, the uniqueness theorem – see e.g. Theorem 3.18 in [17] – ensures that any other measurement M' on that structure is a linear positive transformation of M, that is $\mathrm{M}' = \alpha\mathrm{M} + \beta$, $\alpha, \beta \in \mathbb{R}$.

However, in the case of IR measures, we lack a known interval scale M which we can use to compare all the other IR measures against. Actually, the core issue is even more severe and it is the lack of any notion of interval on the empirical set E of the system runs and, consequently, we cannot define an interval scale M.

Therefore, following [12, 17], we will rely on the notion of *difference structure* 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 $a \prec b$ 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 $d', d'' \in R$ 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 $(R, \preceq_d)$ be a difference structure. Then there exist a measurement scale $\mathrm{M} : E \to \mathbb{R}$ such that for every $a, b, c, d \in E$*

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

This theorem ensures us that, if there is a difference structure on the empirical set E, then there exists an interval scale M. We can then resort to the uniqueness theorem mentioned above and look for a linear positive transformation between this M and any another measurement M' to determine if the latter one is on an interval scale as well.

3 POSET, LATTICE, AND HASSE DIAGRAM

As anticipated in Section 1, we will rely on the notion of *poset* and other constructs to introduce a definition of interval on the empirical set E of the system runs. In this section, following [19], we recall some definitions and results that will be useful afterwards.

A partially ordered set P, **poset** for short, is a set with a partial order $\preceq$ defined on it. 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*.

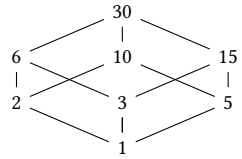
Example. Given a set A, let us consider the set $B = \{E : E \subseteq A\}$ and then define the following ordering: given $E, F \in B$, we say that $E \preceq F$ if $E \subseteq F$. B is the set of all subsets of A *ordered by inclusion* and it is a poset.

Note that a **total order** over a set P is a partial order where every pair of elements are comparable, whereas a **weak order** is a total order without the antisymmetric relation.

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. Note that if $s, t \in P$ lie on the same horizontal level of the diagram, then they are incomparable by construction. Furthermore, elements on different levels may be incomparable as well.

Example. Let $N = 30$ and P the set of all divisors of N, that is $P = \{1, 2, 3, 5, 6, 10, 15, 30\}$. Let us define the following ordering on P: given $a, b \in P$ we say that $a \preceq b$ if a divide b. P is a poset with respect to the ordering $\preceq$, and its Hasse diagram is:



2, 3 and 5 are on the same horizontal level and they are incomparable since, for example, neither 2 divides 3 nor 3 divides 2. Moreover 3 and 10 lie on different levels and they are incomparable.

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. Referring to the previous example, a chain is the subset $\{1, 10, 30\}$, while an example of maximal chain is the subset $\{1, 2, 10, 30\}$.

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. The poset $P = \{1, 2, 3, 5, 6, 10, 15, 30\}$ defined above is graded of rank 3 since any maximal chain of P has length equal to 3.

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

Given $s, t \in P$, an upper bound is $u \in P$ such that $s \preceq u$ and $t \preceq u$. A *least upper bound* (or supremum) of s and t, denoted by $s \vee t$, is an upper bound u such that every other upper bound $v \in P$ of s and t satisfies $v \succeq u$. Dually it is defined the *greatest lower bound* (or infimum) $s \wedge t$. Note that not every pair of elements in a poset has necessarily the infimum or the supremum. A poset L for which every pair of elements has a least upper bound and a greatest lower bound is called **lattice**. The poset P from the previous example is a lattice: for example the elements $2, 15 \in P$ are such that $2 \vee 15 = 30$ and $2 \wedge 15 = 1$, and for any other pair of elements $s, t \in P$ one has $s \vee t$ = *least common multiple* and $s \wedge t$ = *greatest common divisor*.

Proposition 1. *Let L be a finite lattice. The following two conditions are equivalent:*

i. *L is graded, and the rank function ρ of L satisfies*

$$\rho(s) + \rho(t) \geq \rho(s \wedge t) + \rho(s \vee t),$$

for all $s, t \in L$.

ii. *If s and t both covers $s \wedge t$, then $s \vee t$ covers both s and t.*

Moreover, Foldes [8] proves that in a graded poset P the length $\ell(\cdot, \cdot)$ of any interval, also called the **natural distance**, equals the length of the shortest path connecting the two endpoints of the interval in its Hasse diagram.

4 SET-BASED MEASURES

We recall some basic definitions from [7]. Let $(REL, \preceq)$ be a totally ordered set of **relevance degrees** with minimum called the non-relevant relevance degree $\mathrm{nr} = \min(REL)$ and a maximum $\mathrm{rr} = \max(REL)$. In this work, we assume binary relevance, that is we set $REL = \{0, 1\}$ without any loss of generality.

Let us consider a set of **documents** D and a set of **topics** T. For each pair $(t, d) \in T \times D$, the **ground-truth** GT is a map which assigns a relevance degree $rel \in REL$ to a document d with respect to a topic t.

Given a positive natural number N called the *length of the run*, we define the **set of retrieved documents** as $D(N) = \left\{\{d_1, \ldots, d_N\} : d_i \in D\right\}$ and the **universe set of retrieved documents** as $\mathcal{D} := \bigcup_{N=1}^{|D|} D(N)$.

A **run** r_t, retrieving a set of documents $D(N)$ in response to a topic $t \in T$, is a function from T into $\mathcal{D}$

$$t \mapsto r_t = \{d_1, \ldots, d_N\} .$$

A *multiset* (or bag) is a set which may contain the same element several times and its multiplicity of occurrences is relevant [11]. A **set of judged documents** is a multiset $(REL, m) = \{0, 1, 0, \ldots, 0, 0, 1, \ldots\}$, where $m = (m_0, m_1)$ and m_0, m_1 are two functions from REL into $\mathbb{N}_0$ representing the multiplicity of the 0 and 1 relevance degrees, respectively [13]; if the multiplicity is 0, a given relevance degree is not present in the multiset. Let $\mathcal{M}(N)$ be the set of all the possible multiplicity functions m, such that $m_0 + m_1 = N$; then, $\mathcal{R} := \bigcup_{N=1}^{|D|} \bigcup_{m \in \mathcal{M}(N)} (REL, m)$ is the **universe set of judged documents**, i.e. the set of all the possible sets of judged documents (REL, m). 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

$$(t, r_t) \mapsto \hat{r}_t = \left\{GT(t, d_1), \ldots, GT(t, d_N)\right\} = \left\{\hat{r}_{t,1}, \ldots, \hat{r}_{t,N}\right\} .$$

In the following, we omit the dependence on the topic and we simplify the notation into $\hat{r} := \{\hat{r}_1, \ldots, \hat{r}_N\}$, RB, and so on.

As discussed in Section 2.2, we have to start from introducing an order relation $\preceq$ on the set of judged runs. Therefore, we order judged runs with same length by how many relevant documents

they retrieve, i.e. by their total mass of relevance:

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \sum_{i=1}^{N} \hat{r}_i \leq \sum_{i=1}^{N} \hat{s}_i \;. \tag{1}$$

Note that this order is quite intuitive and just corresponds to common sense; therefore it respects the requirement of defining intuitive, sensible and commonly agreeable relations discussed in Section 1.

The order $\preceq$ is a partial order on $\mathcal{R}$, since runs with different length are incomparable. However, to define a difference structure on $\mathcal{R}$ and apply Theorem 1, we need a weak order, that is a totally ordered subset of $\mathcal{R}$ since the antisymmetric relation is satisfied on $\mathcal{R}$.

Let us define $\mathcal{R}(N) := \bigcup_{m \in \mathcal{M}(N)} (REL, m)$ as the set of the judged runs with length fixed to N. $\mathcal{R}(N) \subseteq \mathcal{R}$ and it is a totally ordered set with respect to the ordering $\preceq$ defined in (1) since every pair of runs on this set is comparable. Moreover, $\mathcal{R}(N)$ is a maximal chain of $\mathcal{R}$ since runs with same length are all and only the comparable runs.

Since $\mathcal{R}(N)$ is a totally ordered set and $|\mathcal{R}(N)| = N + 1$, it is *graded of rank N*. Therefore, as discussed in Section 3, there is a unique *rank function* $\rho : \mathcal{R}(N) \to \{0, 1, \ldots, N\}$ which is given by $\rho(\hat{r}) = \sum_{i=1}^{N} \hat{r}_i$. Indeed, $\rho(\{0, \ldots, 0\}) = 0$; if $\hat{s}$ cover $\hat{r}$, that is $\hat{s}$ has one more relevant document than $\hat{r}$, then $\rho(\hat{s}) = \rho(\hat{r}) + 1$ by definition of rank function. This leads for any $r \preceq s$ to the following *natural distance* on $\mathcal{R}(N)$: $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r}) = \sum_{i=1}^{N} (\hat{s}_i - \hat{r}_i)$.

We can finally rely on this natural distance to introduce the definition of difference Δ, which is the core building block for an interval scale as explained in Section 2.2.

Definition 2. Given two runs $\hat{r}, \hat{s} \in \mathcal{R}(N)$, the **difference** between $\hat{r}$ and $\hat{s}$ is defined as $\Delta_{\hat{r}\hat{s}} = \sum_{i=1}^{N} \left(\hat{s}_i - \hat{r}_i\right)$, that is $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

Let $\preceq_d$ be the *less than or equal to* relation on $\mathcal{R}(N) \times \mathcal{R}(N)$, where the subscript d is to highlight its connection with intervals as described in Section 2.2; note that $\preceq_d$ is exactly the order relation $\leq$ among real numbers. We show that $(\mathcal{R}(N), \preceq_d)$ is a *difference structure*. Indeed the first three axioms of Theorem 1 follow immediately from the fact that the ordering $\preceq_d$ between intervals is given by the well known order $\leq$, thanks to the definition of difference. Whereas the *Solvability Condition*, i.e. having an equally-spaced gradation on $\mathcal{R}(N)$, is satisfied by construction: if $\hat{s}$ covers $\hat{r}$, the difference $\Delta_{\hat{r}\hat{s}}$ is constant and equal to 1 ($\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r}) = \rho(\hat{r}) + 1 - \rho(\hat{r}) = 1$).

Let us show how we can construct an interval scale measure M on $\mathcal{R}(N)$. The rank function ρ counts the number of relevant retrieved documents and, if $\hat{s}$ covers $\hat{r}$, the difference $\Delta_{\hat{r}\hat{s}} = \rho(\hat{s}) - \rho(\hat{r})$ is always equal to 1, by construction. Thus an interval scale measure M on $(\mathcal{R}(N), \preceq_d)$ is given by the rank function itself:

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N} \hat{r}_i \;,$$

which satisfies the condition imposed by Theorem 1: let $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in \mathcal{R}(N)$ such that $\Delta_{\hat{r}\hat{s}} \leq \Delta_{\hat{u}\hat{v}}$, then $\Delta_{\hat{r}\hat{s}} \leq \Delta_{\hat{u}\hat{v}} \Leftrightarrow \sum_{i=1}^{N} \left(\hat{s}_i - \hat{r}_i\right) \leq \sum_{i=1}^{N} \left(\hat{v}_i - \hat{u}_i\right) \Leftrightarrow \mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) \leq \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$; thus M is an interval scale on $(\mathcal{R}(N), \preceq_d)$.

We can finally proceed with the last step of Section 2.2 and check whether an IR measure uses an interval scale on $(\mathcal{R}(N), \preceq_d)$ by looking for a linear positive transformation with M.

Let us consider *Precision*

$$\mathrm{Prec@}[N](\hat{r}) = \frac{1}{N} \sum_{i=1}^{N} \hat{r}_i = \frac{\mathrm{M}(\hat{r})}{N} \;;$$

thus Precision is an interval scale.

Similarly, *Recall*

$$\mathrm{Recall}(\hat{r}) = \frac{1}{RB} \sum_{i=1}^{N} \hat{r}_i = \frac{\mathrm{M}(\hat{r})}{RB}$$

is an interval scale.

The *F-measure*, that is the harmonic mean of Precision and Recall,

$$\mathrm{F}(\hat{r}) = 2\frac{\mathrm{Prec}(\hat{r}) \cdot \mathrm{Recall}(\hat{r})}{\mathrm{Prec}(\hat{r}) + \mathrm{Recall}(\hat{r})} = \frac{2}{N + RB} \sum_{i=1}^{N} \hat{r}_i = \frac{2\mathrm{M}(\hat{r})}{N + RB}$$

is an interval scale as well.

4.1 Related Work

van Rijsbergen [21] exploited conjoint structures to study Precision and Recall by considering all the possible Precision and Recall pairs, i.e. $R \times P$, as the empirical set E and then creating a kind of "second order" measure on this set E whose properties are examined, e.g. if this "second order" measure is interval based. We take a different approach since we consider system runs as the empirical set E and not the set of all the possible Precision and Recall pairs; moreover, we directly determine if an IR measure is on an interval scale by exploiting the ordering and difference among system runs.

Bollmann and Cherniavsky [4] introduced the *MZ-metric* and, following the example of van Rijsbergen [21], they defined a conjoint structure on the contingency table relevant/not relevant and retrieved/not retrieved in order to determine under which transformations the MZ-metric was on an interval scale. Instead of a conjoint structure on the contingency table, we directly created a difference structure on the set of system runs that can be used to determine if any set-based IR measure is on an interval scale.

Moreover, the MZ-metric is not on a interval scale if we use the structure we defined above. Indeed, if MZ is the *MZ-metric*, let us consider the measure $\overline{MZ} := 1 - MZ$, since we are working with effectiveness measures, defined as $\overline{MZ}(\hat{r}) = \dfrac{\sum_{i=1}^{N} \hat{r}_i}{RB + N - \sum_{i=1}^{N} \hat{r}_i}$ for $\hat{r} \in \mathcal{R}(N)$. This measure is not interval scale on $(\mathcal{R}(N), \preceq_d)$, since replacing a non relevant document with a relevant one yields to a not constant increment of the measure which depends on the run at hand. This further stresses that the core issue in determining what are the properties of a measure is to agree on what are the appropriate relations among the entities in the empirical set E, from which the properties of a measure are then derived.

Finally, Bollmann [3] studied set-based measures by showing that measures complying with a monotonicity and an Archimedean axiom are a linear combination of the number of relevant retrieved documents and the number of not relevant not retrieved documents. We address a completely different issue, that is determining which scales are used by IR measures.

5 RANK-BASED MEASURES

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.

5.1 Strong Top-Heaviness

Top-heaviness is a central property in IR, stating that the higher a system ranks relevant documents the better it is. If we apply this property at each rank position (not only at the first ones) and we take to extremes the importance of having a relevant document ranked higher, we can define a **strong top-heaviness** property which, in turn, will induce total ordering among runs with fixed length N.

We start from the definition of an order among system runs. Let $\hat{r}, \hat{s} \in REL^N$ such that $\hat{r} \neq \hat{s}$, then there exists $k = \min\{j \leq N : \hat{r}[j] \neq \hat{s}[j]\} < \infty$, and we order system runs as follows

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \hat{r}[k] \leq \hat{s}[k] \,. \tag{2}$$

This ordering prefers a single relevant document ranked higher to any number of relevant documents ranked just below it; more formally, $(\hat{u}[1], \ldots, \hat{u}[m], \mathbf{1}, 0, \ldots, 0)$ is greater than $(\hat{u}[1], \ldots, \hat{u}[m], \mathbf{0}, 1, \ldots, 1)$, for any length $N \in \mathbb{N}$ and for any $m \in \{0, 1, \ldots, N-1\}$. This is why we call it *strong top-heaviness*. This ordering makes sense and it is quite intuitive but it might be considered too radical; therefore, it is a matter of future discussion to determine if it can also be commonly agreed on.

REL^N is totally ordered with respect to $\preceq$, since for every pair of runs $\hat{r}, \hat{s} \in REL^N$, if k is the smallest depth at witch the two runs differ, we establish which one is the biggest by just looking at the values of $\hat{r}[k]$ and $\hat{s}[k]$.

Moreover, REL^N is *graded of rank* $2^N - 1$ since $|\{0, 1\}^N| = 2^N$ and $REL^N = \{0, 1\}^N$ is a maximal chain. Therefore, there is a unique rank function $\rho : REL^N \longrightarrow \{0, 1, \ldots, 2^N - 1\}$ which is given by:

$$\rho(\hat{r}) = \sum_{i=1}^{N} 2^{N-i} \hat{r}[i] \,.$$

If we look at the runs as binary strings, the rank function is exactly the representation in base 10 of the number identified by a run and the ordering among runs $\preceq$ corresponds to the ordering $\leq$ among binary numbers.

Example. Let $\hat{r}, \hat{s} \in REL^5$ be such that $\hat{r} = (0, 0, 1, 1, 1)$ and $\hat{s} = (0, 1, 0, 0, 0)$. Since $\hat{r}[1] = \hat{s}[1]$, while $\hat{r}[2] = 0 < 1 = \hat{s}[2]$, we have $\hat{r} \prec \hat{s}$. Moreover $\rho(\hat{r}) = 2^2 + 2^1 + 2^0 = 7 < 8 = 2^3 = \rho(\hat{s})$ and, in particular, $\hat{s}$ covers $\hat{r}$ (indeed $\rho(\hat{s}) = \rho(\hat{r}) + 1$).

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in REL^N$ such that $\hat{r} \preceq \hat{s}$, and we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

Definition 3. Given two runs $\hat{r}, \hat{s} \in REL^N$, the **difference** between $\hat{r}$ and $\hat{s}$ is defined as $\Delta_{\hat{r}\hat{s}} = \sum_{i=1}^{N} 2^{N-i}\big(\hat{s}[i] - \hat{r}[i]\big)$.

Let $\preceq_d$ be the *less than or equal to* relation on $REL^N \times REL^N$, which as in the set-based case is exactly the order relation $\leq$ among real numbers, then $(REL^N, \preceq_d)$ is a difference structure. Indeed, as shown for the set-based case, the first three axioms of Theorem 1 follow immediately from the fact that the ordering $\preceq_d$ between intervals is given by the well known order $\leq$, thanks to the definition of difference. Finally, the *Solvability Condition*, that is needed to have an equally-spaced gradation on REL^N, is satisfied by construction of the rank function, since $\Delta_{\hat{r}\hat{s}} = \rho(\hat{s}) - \rho(\hat{r}) = 1$ for every $\hat{r}, \hat{s} \in REL^N$ such that $\hat{s}$ covers $\hat{r}$.

Similarly to the set-based case, an interval scale measure M on $(REL^N, \preceq_d)$ is given by the rank function itself

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N} 2^{N-i} \hat{r}[i]$$

which is an interval scale since it satisfies the condition imposed by Theorem 1. To prove it, let $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in REL^N$ such that $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{u}\hat{v}}$; then, $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{u}\hat{v}} \Leftrightarrow \sum_{i=1}^{N} 2^{N-i}\big(\hat{s}[i] - \hat{r}[i]\big) \leq \sum_{i=1}^{N} 2^{N-i}\big(\hat{v}[i] - \hat{u}[i]\big) \Leftrightarrow \mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) \leq \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$, as we have to show.

Remember that a measure M′ is an ordinal scale on REL^N if, for every $\hat{r}, \hat{s} \in REL^N$, the following statement is true:

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \mathrm{M}'(\hat{r}) \leq \mathrm{M}'(\hat{s}) \,.$$

Let us show that RBP_p is ordinal scale on REL^N, with respect to the total ordering defined above, if and only if $p \leq 1/2$. Even though we work with N fixed, note that we want RBP_p ordinal for some $p \geq 0$ to hold regardless of the chosen value for N.

Let us consider $\hat{r}, \hat{s} \in REL^N$ such that $\hat{r} \prec \hat{s}$. Then there exist $k \in \{1, \ldots, N\}$ such that $\hat{r} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathbf{0}, \hat{r}[k+1], \ldots, \hat{r}[N])$ and $\hat{s} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathbf{1}, \hat{s}[k+1], \ldots, \hat{s}[N])$. Let moreover $\hat{\hat{r}}, \hat{\hat{s}} \in REL^N$ be such that $\hat{\hat{r}} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathbf{0}, 1, \ldots, 1)$ and $\hat{\hat{s}} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathbf{1}, 0, \ldots, 0)$. Clearly $\hat{r} \preceq \hat{\hat{r}} \prec \hat{\hat{s}} \preceq \hat{s}$, let us prove that $\mathrm{RBP}_p(\hat{\hat{r}}) < \mathrm{RBP}_p(\hat{\hat{s}})$ iff $p \leq 1/2$.

$\mathrm{RBP}_p(\hat{\hat{r}}) - \mathrm{RBP}_p(\hat{\hat{s}}) = (1-p)\left(\sum_{i=1}^{k-1} \hat{r}[i]p^{i-1} + \sum_{i=k+1}^{N} p^{i-1}\right) - (1-p)\left(\sum_{i=1}^{k-1} \hat{r}[i]p^{i-1} + p^{k-1}\right) = (1-p)\left(\sum_{i=k+1}^{N} p^{i-1} - p^{k-1}\right) = (1-p)\left(\dfrac{p^k - p^N}{1-p} - p^{k-1}\right) = -p^{k-1}(1 - 2p + p^{N-k+1})$.

Note that $\text{RBP}_p(\hat{\bar{r}}) < \text{RBP}_p(\hat{\bar{s}}) \Leftrightarrow \text{RBP}_p(\hat{\bar{r}}) - \text{RBP}_p(\hat{\bar{s}}) < 0 \Leftrightarrow 1 - 2p + p^{N-k+1} > 0$. If $p > 1/2$ then there exists $N \in \mathbb{N}$ big enough such that $1 - 2p + p^{N-k+1} < 0$, while if $p \leq 1/2$ then $1 - 2p \geq 0$ and it is true that $1 - 2p + p^{N-k+1} > 0$. Then we have shown that $\text{RBP}_p(\hat{\bar{r}}) < \text{RBP}_p(\hat{\bar{s}}) \Leftrightarrow p \leq 1/2$.

Moreover $\text{RBP}_p(\hat{r}) \leq \text{RBP}_p(\hat{\bar{r}})$, since $\hat{r}[i] \leq 1$ for all $i \in \{1, \dots, N\}$, and $\text{RBP}_p(\hat{\bar{s}}) \leq \text{RBP}_p(\hat{s})$. Thus we can conclude that, for $p \leq 1/2$, $\hat{r} \prec \hat{s} \Rightarrow \text{RBP}_p(\hat{r}) < \text{RBP}_p(\hat{s})$. Moreover, if $\hat{r} = \hat{s}$, then simply $\text{RBP}_p(\hat{r}) = \text{RBP}_p(\hat{s})$. Therefore, for $p \leq 1/2$, $\hat{r} \preceq \hat{s} \Rightarrow \text{RBP}_p(\hat{r}) \leq \text{RBP}_p(\hat{s})$.

To show the other implication of the *iff*, that is $\text{RBP}_p(\hat{r}) \leq \text{RBP}_p(\hat{s}) \Rightarrow \hat{r} \preceq \hat{s}$, we just prove that $\text{not}\{\hat{r} \preceq \hat{s}\} \Rightarrow \text{not}\{\text{RBP}_p(\hat{r}) \leq \text{RBP}_p(\hat{s})\}$, i.e. we need to prove that $\hat{r} \succ \hat{s} \Rightarrow \text{RBP}_p(\hat{r}) > \text{RBP}_p(\hat{s})$. But this last relation is exactly what we have already proven above, exchanging $\hat{r}$ and $\hat{s}$, hence the proof is complete.

RBP_p with $p > 1/2$ and other IR measures – namely DCG, AP and ERR – are not even ordinal scale on REL^N. Indeed, let us for example consider the runs $\hat{r} = (0, 0, 1, 1, 1)$ and $\hat{s} = (0, 1, 0, 0, 0)$ on REL^5: clearly $\hat{r} \preceq \hat{s}$. Note instead that $\text{DCG}_2(\hat{r}) = 1/\log_2 3 + 1/\log_2 4 + 1/\log_2 5 > 1 = \text{DCG}_2(\hat{s})$; $RB{\cdot}\text{AP}(\hat{r}) = 1/3 + 2/4 + 3/5 > 1/2 = RB{\cdot}\text{AP}(\hat{s})$, where RB is the recall base; $\text{ERR}(\hat{r}) = 1/6 + 1/16 + 1/40 > 1/4 = \text{ERR}(\hat{s})$; finally, $\text{RBP}_p(\hat{r}) = (1-p)(p^2 + p^3 + p^4) > (1-p)p = \text{RBP}(\hat{s})$ for $p \gtrsim 0.54$, and such an example can be found for any other values of $p > 1/2$. Hence, these measures cannot be on interval scale, since an interval scale measure is also ordinal scale.

Therefore, only RBP_p with $p \leq 1/2$ may be on interval scale. Note that only $\text{RBP}_{1/2}$ is a linear positive transformation of the M defined above:

$$\text{RBP}_{1/2}(\hat{r}) = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{2^{i-1}} \hat{r}[i] = \frac{1}{2^N} \sum_{i=1}^{N} 2^{N-i} \hat{r}[i] = \frac{1}{2^N} \text{M}(\hat{r}) \,,$$

for every $\hat{r} \in REL^N$. While RBP_p with $p < 1/2$ is not a linear positive transformation of M, since it does not preserve the equivalence between differences. Indeed, let us consider $\hat{r} = (0, 0, 0, 0, 1)$, $\hat{s} = (0, 0, 0, 1, 0)$, $\hat{u} = (0, 0, 0, 1, 1)$ and $\hat{v} = (0, 0, 1, 0, 0)$, four runs on REL^5. Note that $\hat{s}$ covers $\hat{r}$ and $\hat{v}$ covers $\hat{u}$, but $\text{RBP}_p(\hat{s}) - \text{RBP}_p(\hat{r}) = \text{RBP}_p(\hat{v}) - \text{RBP}_p(\hat{u}) \Leftrightarrow (1-p)(p^3 - p^4) = (1-p)(p^2 - p^3 - p^4)$, that is *iff* $p = 1/2$, as we expect.

Therefore we have shown that, given the total order (2) induced by the strong top-heaviness, $\text{RBP}_{1/2}$ is the only one among the considered IR measures that is on an interval scale with respect to the difference structure defined above.

5.2 Weak Top-Heaviness

In this section, we abandon the total ordering induced by the strong top-heaviness and we explore another ordering, induced by a weaker form of top-heaviness. This ordering is based on these two *monotonicity-like* properties proposed by Ferrante et al. [7]:

- **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

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \sum_{j=1}^{k} \hat{r}[j] \leq \sum_{j=1}^{k} \hat{s}[j] \;\; \forall k \in \{1, \dots, N\} \,. \tag{3}$$

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. With respect to the strong top-heaviness of eq. (2), this ordering is less extreme because it is sensitive to the total mass of relevance accumulated at the different rank positions instead of "cutting" everything just because of a single relevant document ranked higher. This is why we call it *weak top-heaviness*. Moreover, this ordering is based on two monotonicity-like properties which are common-sense and have been somehow pointed out also in other previous works [2, 14]. Therefore, being also intuitive and sensitive, this ordering might be commonly agreed on in an easier way than the strong top-heaviness ordering.

The ordering $\preceq$ is a partial ordering on REL^N: for example, when $N = 5$ the runs $\hat{r} = (0, 1, 1, 0, 1)$ and $\hat{s} = (1, 0, 0, 0, 1)$ are incomparable, since $\hat{s}[1] > \hat{r}[1]$ while $\sum_{i=1}^{3} \hat{s}[i] = 1 < 2 = \sum_{i=1}^{3} \hat{r}[i]$. Thus REL^N is a poset. In addition REL^N is a lattice; indeed, for every $\hat{r}, \hat{s} \in REL^N$, $\hat{r} \wedge \hat{s} \geq (0, \dots, 0)$ and $\hat{r} \vee \hat{s} \leq (1, \dots, 1)$.

Since REL^N is a poset, that is it does not have a weak order, we have no chance to find a difference structure defined on the whole set, as we did in Section 5.1. Thus we first have to highlight some properties associated to REL^N as a poset, and then we will make use of totally ordered subsets of REL^N, i.e. chains, where it is possible to define a difference structure.

Proposition 2. *Let $N \in \mathbb{N}$ be fixed and $REL = \{0, 1\}$. The poset REL^N is graded, i.e. every maximal chain of REL^N has the same length.*

Proof. Thanks to Proposition 1 and since REL^N is a lattice, it is sufficient to prove that for each $\hat{r}, \hat{s} \in REL^N$ that both cover $\hat{r} \wedge \hat{s}$, $\hat{r} \vee \hat{s}$ covers both $\hat{r}$ and $\hat{s}$.

Let $\hat{r}, \hat{s} \in REL^N$ be such that $\hat{r} \prec \hat{s}$, define $c = \left|\{k \leq N : \sum_{i=1}^{N} \hat{r}[i] < \sum_{i=1}^{N} \hat{s}[i]\}\right|$ and denote with $k_1 < \dots < k_c$ the depths where the strict inequality on (3) hold. Firstly note that if $\hat{r} \prec \hat{s}$ then $c \geq 1$.

If $c = 1$ and $k_1 < N$, then $\hat{s}$ and $\hat{r}$ differ in a **swap** of length one $\hat{s} = (\dots, \hat{s}[k_1 - 1], \mathbf{1}, \mathbf{0}, \hat{s}[k_1 + 2], \dots)$, $\hat{r} = (\dots, \hat{s}[k_1 - 1], \mathbf{0}, \mathbf{1}, \hat{s}[k_1 + 2], \dots)$. If $c = 1$ and $k_1 = N$, then $\hat{s}$ and $\hat{r}$ differ in a **replacement** in the last position: $\hat{s} = (\dots, \hat{s}[k_1 - 1], \mathbf{1})$, $\hat{r} = (\dots, \hat{s}[k_1 - 1], \mathbf{0})$.

In both cases, for every $\hat{u} \in REL^N$ such that $\hat{r} \preceq \hat{u} \preceq \hat{s}$, then $\hat{u} = \hat{r}$ or $\hat{u} = \hat{s}$, and this follows immediately from the partial order recalled above.

On the contrary, if $c > 1$, there are two cases to study: $k_2 > k_1 + 1$ or $k_2 = k_1 + 1$. In the first case we have the following situation:
$\hat{s} = (\dots, \hat{s}[k_1 - 1], \mathbf{1}, \mathbf{0}, \hat{s}[k_1 + 2], \dots, \hat{s}[k_2 - 1], \mathbf{1}, \hat{s}[k_2 + 1], \dots)$,
$\hat{r} = (\dots, \hat{s}[k_1 - 1], \mathbf{0}, \mathbf{1}, \hat{s}[k_1 + 2], \dots, \hat{s}[k_2 - 1], \mathbf{0}, \hat{r}[k_2 + 1], \dots)$.
Note that $\hat{r}[k_1 + 1] = 1$ while $\hat{s}[k_1 + 1] = 0$ since $\sum_{i=1}^{k_1+1} \hat{r}[i]$ has to be equal to $\sum_{i=1}^{k_1+1} \hat{s}[i]$ as $k_2 > k_1 + 1$. Then the following run
$\hat{u} = (\dots, \hat{s}[k_1 - 1], \mathbf{0}, \mathbf{1}, \hat{s}[k_1 + 2], \dots, \hat{s}[k_2 - 1], \mathbf{1}, \hat{s}[k_2 + 1], \dots)$
is such that $\hat{r} \prec \hat{u} \prec \hat{s}$.

While when $k_2 = k_1 + 1$, we can have $\hat{s} = (\dots, \hat{s}[k_1 - 1], \mathbf{1}, \mathbf{0}, \hat{s}[k_2 + 1], \dots)$ and $\hat{r} = (\dots, \hat{s}[k_1 - 1], \mathbf{0}, \mathbf{0}, \hat{r}[k_2 + 1], \dots)$,

or $\hat{s} = (\dots, \hat{s}[k_1 - 1], \mathbf{1}, \mathbf{1}, \hat{s}[k_2 + 1], \dots)$ and $\hat{r} = (\dots, \hat{s}[k_1 - 1], \mathbf{0}, \mathbf{0}, \hat{r}[k_2 + 1], \dots)$. In both cases, $\hat{u}$ given by $\hat{u} = (\dots, \hat{s}[k_1 - 1], \mathbf{0}, \mathbf{1}, \hat{s}[k_2 + 1], \dots)$ is such that $\hat{r} \prec \hat{u} \prec \hat{s}$.

Thus we have shown that the "cover" relations, that is the operation for which from a run we can obtain a new run that covers the first one, are swap of length one and replacements in the last position.

Now let $\hat{r}, \hat{s} \in \{0,1\}^N$ such that both cover $\hat{r} \wedge \hat{s}$, which implies that $\hat{r}$ and $\hat{s}$ are incomparable. This means that does not exists $\hat{z} \in REL^N$ such that $\hat{r} \wedge \hat{s} \prec \hat{z} \prec \hat{r}$ nor $\hat{r} \wedge \hat{s} \prec \hat{z} \prec \hat{s}$. Thus, if $\hat{u} := \hat{r} \wedge \hat{s} := (\hat{u}[1], \dots, \hat{u}[N])$, there exist an index $i \in \{1, \dots, N-1\}$ such that $u[i] = 0$ and $u[i+1] = 1$. Since $\hat{r}$ and $\hat{s}$ both cover $\hat{u}$, we have two possibilities (up to symmetries):

i. $\hat{r} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathbf{1}, \mathbf{0}, \hat{u}[i+2], \dots, \hat{u}[N])$ and $\hat{s} = (\hat{u}[1], \dots, \hat{u}[j-1], \mathbf{1}, \mathbf{0}, \hat{u}[j+2], \dots, \hat{u}[N])$, if $\hat{u}[j] = 0$, $\hat{u}[j+1] = 1$, where $j > i+1$;
ii. $\hat{r} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathbf{1}, \mathbf{0}, \hat{u}[i+2], \dots, \hat{u}[N])$ and $\hat{s} = (\hat{u}[1], \dots, \hat{u}[N-1], \mathbf{1})$, if $\hat{u}[N] = 0$.

Respectively, let us define $\hat{t} \in REL^N$ as

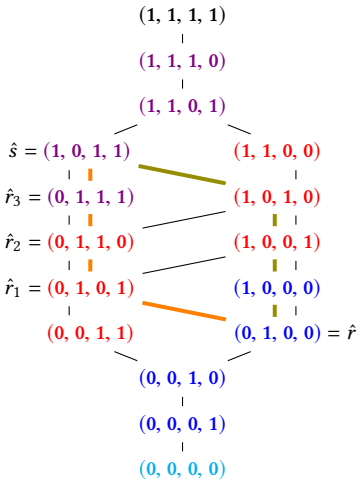
i. $\hat{t} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathbf{1}, \mathbf{0}, \hat{u}[i+2], \dots, \hat{u}[j-1], \mathbf{1}, \mathbf{0}, \hat{u}[j+2], \dots, \hat{u}[N])$;
ii. $\hat{t} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathbf{1}, \mathbf{0}, \hat{u}[i+2], \dots, \hat{u}[N-1], \mathbf{1})$.

The notes made above entail that $\hat{t}$ covers both $\hat{r}$ and $\hat{s}$, since $\hat{t}$ differs from each of them only for one swap or a replacement in the last position. Then $\hat{t} = \hat{r} \vee \hat{s}$ and the proof is complete. □

As discussed in Section 3, REL^N graded implies that the natural distance $\ell(\cdot,\cdot)$ is well defined for every two comparable elements $\hat{r}, \hat{s} \in REL^N$ as the length of a maximal chain in $[\hat{r}, \hat{s}]$ minus 1. Equivalently, given the *rank function* $\rho : REL^N \longrightarrow \mathbb{N}$, the natural distance is defined as $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$ and, if $\hat{s}$ covers $\hat{r}$, then $\rho(\hat{s}) = \rho(\hat{r}) + 1$.

Remember that the natural length of any interval of a graded poset equals the numbers of edges in every shortest path connecting the endpoints of the interval in its Hasse diagram. The next example highlights this fact.

Example. Let us fix $N = 4$. The Hasse Diagram of REL^N is



where different colours of the runs correspond to different total numbers of relevant retrieved documents.

Given $\hat{r} = (0,1,0,0)$, $\hat{s} = (1,0,1,1)$, let us consider one of the shortest paths between the two runs, for example the one with orange edges: it starts from $\hat{r}$, goes through $\hat{r}_1, \hat{r}_2, \hat{r}_3$, and ends in $\hat{s}$. Note that $\{\hat{r}_0 := \hat{r}, \hat{r}_1, \hat{r}_2, \hat{r}_3, \hat{r}_4 := \hat{s}\}$ is also a maximal chain, since there does not exist $\hat{u} \in REL^4$ such that $\hat{r}_i \prec \hat{u} \prec \hat{r}_{i+1}$ for some $i \in \{0,1,2,3\}$. Moreover this shortest path has length 4, and every other shortest path between $\hat{r}$ and $\hat{s}$ has the same length, e.g. the one with dark green edges. Thus the natural length of $[\hat{r}, \hat{s}]$ is 4.

The explicit expression for the rank function is

$$\rho(\hat{r}) = \sum_{i=1}^{N} (N - i + 1)\hat{r}[i] \,.$$

Indeed, recalling that given two runs one covers the other if they differ only for a swap of length one or a replacement in the last position, in order to compute $\rho(\hat{r})$ we need to count the number of replacements and swaps needed to go from the smallest run possible, i.e. $(0, \dots, 0)$ to $\hat{r}$ along a path in the Hasse diagram, where the edges are the "cover" relations.

Example. Let us consider $\hat{s} = (1,0,1,1)$ from the previous example. Since $\hat{s}[1] = 1$, from $\hat{0} = (0,0,0,0)$ we need a replacement in $\hat{0}[4]$ plus three swaps to reach $\hat{s}[1]$, that is we have to do four "cover" operation to go from $\hat{0}$ to $(1,0,0,0)$ and, equivalently, the path in the Hasse diagram has length equal to 4. Since $\hat{s}[3] = 1$, from $(1,0,0,0)$ to $(1,0,1,0)$ we need a replacement in the last position plus a swap, that is 2 more "cover" operations. Eventually, with another replacement, we reach $\hat{s}$. Hence $\rho(\hat{s}) = 4 + 2 + 1 = 7 = \sum_{i=1}^{4}(4 - i + 1)\hat{s}[i]$, as stated.

Therefore, from the natural distance and the rank function, we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

Definition 4. Given two comparable runs $\hat{r}, \hat{s} \in REL^N$, the **difference** between $\hat{r}$ and $\hat{s}$ is $\Delta_{\hat{r}\hat{s}} = \sum_{i=1}^{N}(N - i + 1)\big(\hat{s}[i] - \hat{r}[i]\big)$.

Note that, contrary to the previous cases, since the ordering given by (3) is only partial, in order to compare differences between intervals we need to restrict our study to a maximal chain, i.e. a totally ordered subset of REL^N. Thus, denoted with $C(REL^N)$ a maximal chain of REL^N, and given the *less than or equal to* $\preceq_d$ relation, which as in the previous cases coincides with the order relation $\leq$ among real numbers, the relational structure $(C(REL^N), \preceq_d)$ is a difference structure. This follows from the same discussion we have done for the difference structure in the strong top-heaviness case in the previous section.

Therefore, an interval scale measure M on $(C(REL^N), \preceq_d)$ is given by the rank function, that is

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N}(N - i + 1)\hat{r}[i] \,,$$

for $\hat{r} \in C(REL^N)$.

AP, RBP, DCG, and ERR are on a ordinal scale with respect the partial ordering (3) induced by the weak top-heaviness, as demonstrated by [7]. However, none of them is on an interval scale since there does not exist any positive linear transformation between M and any of them. In particular, the next example shows how each of them fails on intervals with same length.

Example. Consider the following runs on REL^4 : $\hat{r} = (0, \mathbf{1}, 0, 0)$, $\hat{s} = (\mathbf{1}, 0, 0, 0)$, $\hat{u} = (0, 0, 0, \mathbf{1})$ and $\hat{v} = (0, 0, \mathbf{1}, 0)$. These runs are comparable, that is they belong to the same maximal chain on REL^4. Moreover, $\hat{s}$ covers $\hat{r}$ and $\hat{v}$ covers $\hat{u}$, that is the differences $\Delta_{\hat{r}\hat{s}}$ and $\Delta_{\hat{u}\hat{v}}$ are equal.

Hence an interval scale measure M should satisfy $\mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) = \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$, as a consequence of Theorem 1. However, in the case of AP we have that $RB \cdot (\mathrm{AP}(\hat{s}) - \mathrm{AP}(\hat{r})) = 1 - 1/2 > 1/3 - 1/4 = (\mathrm{AP}(\hat{v}) - \mathrm{AP}(\hat{u})) \cdot RB$, where RB is the recall base. In the case of RBP we have that $\mathrm{RBP}_p(\hat{s}) - \mathrm{RBP}_p(\hat{r}) = (1-p)^2 > (1-p)^2 p^2 = \mathrm{RBP}_p(\hat{v}) - \mathrm{RBP}_p(\hat{u})$ since $p < 1$. In the case of DCG we have that $\mathrm{DCG}_2(\hat{s}) - \mathrm{DCG}_2(\hat{r}) = 1 - 1 < 1/\log_2 3 - 1/\log_2 4 = \mathrm{DCG}_2(\hat{v}) - \mathrm{DCG}_2(\hat{u})$. Finally, in the case of ERR we have that $\mathrm{ERR}(\hat{s}) - \mathrm{ERR}(\hat{r}) = 1/2 - 1/4 > 1/6 - 1/8 = \mathrm{ERR}(\hat{v}) - \mathrm{ERR}(\hat{u})$. This proves that none of these measures is an interval scale on $(C(REL^N), \preceq_d)$, where $C(REL^N)$ is such that $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in C(REL^N)$.

5.3 Related Work

Both Amigó et al. [2] and Moffat [14] studied the properties of IR measures, in a formal and a numeric way respectively, defining, e.g., how an IR measure should behave when a relevant document is added or removed from a system run. All the identified properties could be exploited to introduce some sort of structure among the system runs but these authors did not do that explicitly. Moreover, they did not study what scales are adopted by IR measures, which is the core topic of this paper instead.

Busin and Mizzaro [5] used the notion of scale and mapping among scale to model different kinds of similarity and to introduce constraints and axioms over them. However, they did not address the problem of determining the scales used by an IR measure.

6 CONCLUSIONS AND FUTURE WORK

In this paper we have explored the question whether IR evaluation measures are based on an interval scale or not. This is a core issue since the validity of the statistics, such as mean and variance, and the statistical tests we use to compare IR systems depends on the scale adopted by IR measures.

We have relied on the representational theory of measurement and highlighted that the key point to understand the properties of IR measures is to have a clear understanding of the relations among system runs. In particular, to determine if an IR measure is on a interval scale, we need first to have a commonly agreed notion of ordering and a notion of interval among system runs. We have shown how to define such notions of ordering and interval and how to exploit them to determine whether an IR measure is on an interval scale.

In the case of set-based measures and system runs of fixed length, we found that the most popular ones – namely Precision, Recall, and F-measure – are on an interval scale. In the case of rank-based measures and system runs of fixed length, adopting a strongly top-heavy ordering, we found that: RBP with $p = \frac{1}{2}$ is on an interval scale; RBP with $p < \frac{1}{2}$ is on an ordinal scale but not on an interval one; RBP with $p > \frac{1}{2}$, AP, DCG, and ERR are not on an interval scale and not even on an ordinal one. Using a weakly top-heavy ordering, we found that RBP, AP, DCG, ERR are not on an interval scale even if they are on an ordinal one.

Future work will concern further investigation of rank-based measures and we will explore two alternatives. Firstly, instead of defining a notion of ordering among the system runs, we will use the ordering of systems induced by an IR measure itself and we will check if, at least in this case, IR measures are on an interval scale. Secondly, we will relax the properties of Definition 1 by removing the Solvability Condition. This will cause the intervals of system runs to not be anymore equi-spaced but could allow us to introduce a notion of partially interval scale which IR measures might (or not) comply to.

REFERENCES

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

[2] 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.

[3] P. Bollmann. 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Ferrante, N. Ferro, and M. Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness, See [1], 21–30.

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

[9] N. Fuhr. 2012. Salton Award Lecture: Information Retrieval As Engineering Science. *SIGIR Forum* 46, 2 (December 2012), 19–28.

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

[11] D. E. Knuth. 1981. *The Art of Computer Programming – Volume 2: Seminumerical Algorithms* (2nd ed.). Addison-Wesley, USA.

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

[13] S. Miyamoto. 2004. Generalizations of Multisets and Rough Approximations. *International Journal of Intelligent Systems* 19, 7 (July 2004), 639–652.

[14] A. Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Proc. 9th Asia Information Retrieval Societies Conference (AIRS 2013)*, R. E. Banchs, F. Silvestri, T.-Y. Liu, M. Zhang, S. Gao, and J. Lang (Eds.), Vol. 8281. LNCS 8281, Springer, Heidelberg, Germany, 1–12.

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

[16] 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.

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

[18] F. Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms, See [1], 11–20.

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

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

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

[22] P. F. Velleman and L. Wilkinson. 1993. Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician* 47, 1 (February 1993), 65–72.