Information Retrieval Journal (2020) 23:289–317
https://doi.org/10.1007/s10791-019-09362-z

AXIOMATIC THINKING FOR INFORMATION RETRIEVAL

How do interval scales help us with better understanding IR evaluation measures?

Marco Ferrante[1] · **Nicola Ferro**[2] · **Eleonora Losiouk**[1]

Received: 31 July 2018 / Accepted: 28 August 2019 / Published online: 4 September 2019

Abstract
Evaluation measures are the basis for quantifying the performance of IR systems and the way in which their values can be processed to perform statistical analyses depends on the scales on which these measures are defined. For example, mean and variance should be computed only when relying on interval scales. In our previous work we defined a theory of IR evaluation measures, based on the representational theory of measurement, which allowed us to determine whether and when IR measures are interval scales. We found that common set-based retrieval measures—namely precision, recall, and F-measure—always are interval scales in the case of binary relevance while this does not happen in the multi-graded relevance case. In the case of rank-based retrieval measures—namely AP, gRBP, DCG, and ERR—only gRBP is an interval scale when we choose a specific value of the parameter p and define a specific total order among systems while all the other IR measures are not interval scales. In this work, we build on our previous findings and we carry out an extensive evaluation, based on standard TREC collections, to study how our theoretical findings impact on the experimental ones. In particular, we conduct a correlation analysis to study the relationship among the above-mentioned state-of-the-art evaluation measures and their scales. We study how the scales of evaluation measures impact on non parametric and parametric statistical tests for multiple comparisons of IR system performance. Finally, we analyse how incomplete information and pool downsampling affect different scales and evaluation measures.

Keywords Experimentation · Representational theory of measurement · Interval scale · IR evaluation measure · Formal framework

✉ Nicola Ferro
ferro@dei.unipd.it

Marco Ferrante
ferrante@math.unipd.it

Eleonora Losiouk
elosiouk@math.unipd.it

[1] Department of Mathematics "Tullio Levi-Civita", University of Padua, Padua, Italy

[2] Department of Information Engineering, University of Padua, Padua, Italy

1 Introduction

Information Retrieval (IR) faces an extremely challenging task, i.e., ranking typically heterogeneous and very diverse information sources with respect to often vague user information needs for tasks which are more and more demanding and complex. Therefore, even if laying solid foundations has always been a goal of the discipline, the development of formal theories has been always partnered with very systematic and thorough experimentation, needed to assess the performance of IR systems and understand their behaviour.

Recently, there has been a return and a new raise of interest for developing and applying axiomatic methods to IR (Amigó et al. 2017, 2018a) by focusing, for example, on the definition of ranking functions based on axiomatic constraints and on the development of frameworks to model IR evaluation measures. This renewed interest is motivated not only by the push to strengthen the field and lay more rigorous foundations, but also, and mostly, by the need to face new and hard challenges, such as the possibility of predicting the performance of IR systems before developing and experimenting with them (Allan et al. 2018a; Ferro et al. 2018), which calls for better theoretical foundations in the field.

In this context, we have recently developed a general theory of offline IR evaluation measures (Ferrante et al. 2019), which is based on the *representational theory of measurement* adopted in physics (Krantz et al. 1971). *Measurement scales* are central notion in the representational theory of measurement and Stevens (1946) identifies four major types of scales with increasing properties: (1) the *nominal scale* consists of discrete unordered values, i.e. categories; (2) the *ordinal scale* introduces a natural order among the values; (3) the *interval scale* preserves the equality of intervals or differences; and, (4) the *ratio scale* preserves the equality of ratios. Measurement scales are important, since they determine the operations that can be performed with the measured values and, as a consequence, the statistical analyses that can be applied; for example, mean and variance should be computed only if your measurement is an interval scale.

Our theory provides us with a constructive way to define interval scales, in the case of both set-based and rank-based evaluation measures, accommodating both binary and multi-graded relevance judgements. It allows us to formally determine the scale of an evaluation measure and it also introduces new evaluation measures which guarantee to be interval scales.

In this paper, we move a step forward and complement our theory of offline evaluation measures with a thorough experimentation whose overall goal is to explore the effects of meeting or not the assumptions of the scales behind evaluation measures. In particular, we consider several state-of-the-art offline evaluation measures—namely precision, recall, F-measure, AP, RBP, DCG, and ERR—and we compare them to the behaviour of our interval scale measures, namely SBTO in the set-based retrieval case and RBTO in the rank-based retrieval case. We rely on standard TREC collections with both binary and multi-graded relevance judgements to break down our overall goal into the following three research questions:

- RQ1 what is the relationship between the different evaluation measures and how are these affected by their scales?
- RQ2 what is the impact of violating the scale assumptions behind statistical significance tests for comparing IR systems?
- RQ3 how much do less and less complete pools affect evaluation measures and to what extent do the used scales play a role in this?

To the best of our knowledge, this paper represents the first experimental study aimed at quantifying the impact of the scale assumptions behind evaluation measures. Moreover, it also represents the first attempt to turn the findings of a formal theory of IR evaluation measures into an actual experimental validation.

The paper is organized as follows. Section 2 discusses the related works; Sect. 3 reports some background information about set theory and the representational theory of measurement; Sect. 4 briefly summarizes the main points of our theory of IR evaluation measures, only those relevant to the subsequent experimentation; Sect. 5 introduces our experiments and discusses the experimental findings; finally, Sect. 6 wraps up the discussion and outlooks some future work.

2 Related work

The relation between the representational theory of measurements and IR evaluation measures has been early investigated by van Rijsbergen (1974, 1979) in the context of set-based IR measures. In particular, van Rijsbergen (1974) exploited *conjoint structures* (Krantz et al. 1971) to study Precision and Recall.

Bollmann and Cherniavsky (1980) introduced the *MZ-metric* and, following the example of van Rijsbergen , 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. Bollmann (1984) 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.

Amigó et al. (2009, 2013) and Moffat (2013) studied the properties of rank-based IR measures, in a formal and numeric way respectively, defining, e.g., how an IR measure should behave when a relevant document is added or removed from a system run. Recently, Amigó et al. (2018b) extended this approach to diversity-oriented evaluation measures.

Busin and Mizzaro (2013) and Maddalena and Mizzaro (2014) used the notion of scale and mapping among scales to model different kinds of similarity and to introduce constraints over them.

We developed our theory of evaluation measures starting from the exploration of ordinal scales (Ferrante et al. 2015) and then we moved to interval scales in the binary relevance case (Ferrante et al. 2017b). Finally, we consolidated our findings into a single coherent framework and we generalized it to consider also multi-graded relevance and different types of orders among runs (Ferrante et al. 2019). We also started to explore whether it is possible to define semi-interval scales (Ferrante et al. 2017a) and to accommodate IR evaluation measures over them. This paper complements our previous work with the first experimental investigation ever on assessing the impact of scale assumptions and quantifying it from different points of view, i.e. the three research questions. Moreover, it is the first experimental study on evaluation measures motivated and backed by the findings of a formal theory of IR evaluation measures.

3 Background

In this section we summarize some background information about poset, measurement scales, and how to proceed to define interval scales for IR.

3.1 Poset, totally ordered sets, intervals and their length

A partially ordered set P, poset for short, is a set with a partial order $\preceq$ defined on it (Stanley 2012). 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*. P is called bounded if it has a maximum and a minimum element, namely $\hat{1}, \hat{0} \in P$ such that for every $s \in P$, $s \preceq \hat{1}$ and $\hat{0} \preceq s$.

A total order over a poset P is a partial order where every pair of elements are comparable.

A closed interval is a subset of a poset 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 $[s, t] = \{s, t\}$, which means that there is no $u \in P$ such that $s \prec u \prec t$.

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 the order is total, the unique maximal chain is the whole set P.

If every maximal chain of P has the same length n, we say that the poset 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 $\ell(s, t) := \ell([s, t]) = \rho(t) - \rho(s)$.

3.2 Representational theory of measurement

The *representational theory of measurement* (Krantz et al. 1971) sees measurement as the process of assigning numbers to entities in the real world conforming to some property under examination. 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.

Moving to the IR context, being an interval scale is not just a numeric property of an evaluation measure, but firstly we need to understand how system runs are *ordered* among themselves, then what *intervals* of system runs are, and finally how these intervals are ordered too. Only at this point, we can verify whether an evaluation measure complies with these notions and determine whether it is an interval scale.

More precisely, a relational structure (Krantz et al. 1971; Rossi 2014) 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: (1) 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; (2) 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; (3) $\forall r \in R_X, \forall x_i \in X$, 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 entities in the real world, i.e. 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.

3.3 Measurement scales

There are four major types of measurement scales (Stevens 1946) which can be ordered by their increasing properties and allow 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 (Velleman and Wilkinson 1993) 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 (Rossi 2014)—ensures that any other measurement M' on that structure is a linear positive transformation of M, that is $\mathrm{M}' = \alpha\mathrm{M} + \beta$, $\alpha, \beta \in \mathbb{R}$ and $\alpha > 0$.

However, in the case of IR evaluation measures, we lacked a known interval scale M to be used to compare all the other IR measures against. Actually, the core issue was even more severe: we lacked any notion of order on the empirical set E of the IR system runs, thus we also lacked the notion of interval of system runs and, consequently, it was not possible to define an interval scale M too.

In our theory of IR evaluation measures (Ferrante et al. 2019), we overcame these issues by relying on the notion of *difference structure* (Krantz et al. 1971; Rossi 2014) to introduce a definition of interval among system runs and to ensure 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 $b \preceq c \;\Rightarrow\; a \preceq c$ (transitivity).

Note that if $a, b \in E$ are such that $a \preceq b$ and $b \preceq a$, then we write $a \sim b$ and we say that a and b are equivalent elements of E for $\preceq$. This does not necessarily mean that a and b are equal, i.e. $a = b$, since they might be two distinct objects. When the antisymmetric relation holds, that is when $a \preceq b$ and $b \preceq a$ implies that a and b are the same element (namely $a = b$), we talk about a *total order*.

An interval on the empirical structure is an element $(a, b) \in E \times E$ and we introduce a notion of difference Δ_{ab} over intervals, to act as a signed distance we exploit to compare intervals. Once we have a notion of difference Δ_{ab}, we can define a weak order $\preceq_d$ between the Δ_{ab} differences and, consequently, among intervals. We can proceed as follows: if two elements $a, b \in E$ are such that $a \sim b$, 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 $\Delta_{aa} \succ_d \Delta_{ba}$, that is there exist a kind of "zero" and the inverse with respect to this "zero".

The following notion of difference structure allows us to verify whether a measurement is an interval scale or not.

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:

1. $\preceq_d$ is *weak order*;
2. if $\Delta_{ab} \preceq_d \Delta_{cd}$, then $\Delta_{dc} \preceq_d \Delta_{ba}$;
3. *Weak Monotonicity:* if $\Delta_{ab} \preceq_d \Delta_{a'b'}$ and $\Delta_{bc} \preceq_d \Delta_{b'c'}$ then $\Delta_{ac} \preceq_d \Delta_{a'c'}$;
4. *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.

The first condition defines an ordering among intervals while the second one sets a sign for differences. The *Weak Monotonicity* condition gives us a rule to compose adjacent intervals; among other things, it tells us that adding a non-null interval to an interval produces a greater interval. The *Solvability Condition* 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 an interval scale measurement $\mathrm{M} : E \to \mathbb{R}$ 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 over it.

Therefore, to study whether IR measures are interval scales or not, Ferrante et al. (2019) proceeded as follows:

1. Define a total ordering among system runs, which allows us also to introduce the notion of interval among runs;
2. Since this set is graded of a given rank n, there exists a unique rank function ρ which assigns a natural number to each run;
3. Define the length of an interval as the natural distance $\Delta_{ab} := \ell(a, b) := \ell([a, b]) = \rho(b) - \rho(a)$;
4. Check whether the set with the above natural length is a difference structure or not;
5. In this case we have a difference structure and we can define an interval scale M as the rank function ρ itself;
6. We can eventually check whether IR measures are a linear positive transformation of this interval scale M and determine whether they are an interval scale.

4 Formal framework

We summarize here a part of our theory for IR evaluation measures (Ferrante et al. 2019) in order to give the reader an idea of how it works and to better understand the foundations which the subsequent experimental part and research questions are built on. Details,

examples, and proofs are omitted for space reasons and can be found in Ferrante et al. (2019).

We also introduce several state-of-the-art IR evaluation measures, which will be then investigated in the experimentation, to show how they can be expressed within our framework and how you can determine the scales they use.

4.1 Basic formalism

Let $(REL, \preceq)$ be a finite and totally ordered set of relevance degrees. We set $REL = \{\mathscr{a}_0, \mathscr{a}_1, \ldots, \mathscr{a}_c\}$ with $\mathscr{a}_i \prec \mathscr{a}_{i+1}$ for all $i \in \{0, \ldots, c-1\}$; REL has a minimum $\mathscr{a}_0$, called the "not relevant" relevance degree.

Let us consider a finite 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.

Let N be a positive natural number called the *length of the run*. We assume that all the runs have same length N, since this is what typically happens in real evaluation settings when you compare IR systems.

We define $D(N)$ as the set of all the possible N retrieved documents.

A run $r : T \to D(N)$ retrieves N documents belonging to $D(N)$ in response to a topic $t \in T$.

Let $R(N)$ be the set of N judged documents, that is the set of all the N possible combinations of relevance degrees.

We call judged run of length N the function $\hat{r}$ from $T \times D(N)$ into $R(N)$ which assigns a relevance degree to each retrieved document, i.e. a judged run $\hat{r}$ is the application of the ground-truth GT function to each element of the run r.

We define the gain function $g : REL \to \mathbb{R}_+$ as the map that assigns a positive real number to any relevance degree. We set, without loss of generality, $g(\mathscr{a}_0) = 0$ and we require g to be strictly increasing.

We define the indicator function for the relevance degrees as $\delta_{\mathscr{a}}(\mathscr{a}_j) = j \ \ \forall j \in \{0, \ldots, c\}$. Note that $\delta_{\mathscr{a}}$ is a particular gain function.

Given the gain function g, the recall base $RB : T \to \mathbb{R}_+$ is the map defined as $RB(t) = \sum_{j=1}^{|D|} g(GT(t, d_j))$. In the binary relevance case when $c = 1$ and $REL = \{\mathscr{a}_0, \mathscr{a}_1\}$, the gain function usually is $g(\mathscr{a}_1) = \delta_{\mathscr{a}}(\mathscr{a}_1) = 1$ and RB counts the total number of relevant documents for a topic.

An evaluation measure is a function $\mathrm{M} : R(N) \to \mathbb{R}_+$ which maps a judged run $\hat{r}$ into a positive real number which quantifies its effectiveness. Note that most of the evaluation measures are normalized and thus the co-domain is the [0, 1] interval.

In the following, we specialize the above definitions to the case of both set-based and rank-based retrieval.

4.1.1 Set-based retrieval

The set of all the possible unordered N retrieved documents is $D(N) = \big\{\{d_1, \ldots, d_N\} : d_i \in D\big\}$. A run r is given by $r = \{d_1, \ldots, d_N\}$. We denote by r_j the jth element of the set r, i.e. $r_j = d_j$.

A *multiset* (or bag) is a set which may contain the same element several times (Knuth 1981). The set of judged documents is a multiset $(REL, m) = \{\mathscr{a}_1, \mathscr{a}_1, \mathscr{a}_0, \ldots, \mathscr{a}_c, \mathscr{a}_2, \mathscr{a}_c, \ldots\}$, where m is a function from REL into $\mathbb{N}_0$

representing the multiplicity of every relevance degree $\mathscr{a}_j$ (Miyamoto 2004); if the multiplicity is 0, a given relevance degree is not present in the multiset. Let $\mathcal{M}$ be the set of all the possible multiplicity functions m, then $REL(\mathcal{M}) := \bigcup_{m\in\mathcal{M}}(REL, m)$ is the universe set of judged documents, i.e. the set of all the possible sets of judged documents (REL, m). We can define the set of N judged documents as $R(N) := \{\hat{r} \in REL(\mathcal{M}) : |\hat{r}| = N\}$.

Note that, since each judged run in $R(N)$ is an unordered set of N relevance degrees, $R(N)$ consists of all the N combinations of $c + 1 = |REL|$ objects with repetition.

We now introduce the definitions of generalized precision and recall (Kekäläinen and Järvelin 2002), which extend precision and recall to the multi-graded relevance case, and of F-measure.

Generalized Precision (gP) is defined as

$$\mathrm{gP}(\hat{r}) = \frac{1}{N}\sum_{i=1}^{N}\frac{g(\hat{r}_i)}{g(\mathscr{a}_c)}\ ,$$

while *Generalized Recall (gR)* as

$$\mathrm{gR}(\hat{r}) = \frac{1}{RB}\sum_{i=1}^{N}\frac{g(\hat{r}_i)}{g(\mathscr{a}_c)},$$

where $1/g(\mathscr{a}_c)$ is needed to normalize the gain function and RB is recall base. Note that gP coincides with Precision (P) and gR coincides with Recall (R) when binary relevance ($c = 1$) is considered.

The *F-measure* works with binary relevance when $REL = \{\mathscr{a}_0, \mathscr{a}_1\}$ and is the harmonic mean of *Precision (P)* and *Recall (R)* given by

$$\mathrm{F}(\hat{r}) = 2\frac{\mathrm{P}(\hat{r})\cdot\mathrm{R}(\hat{r})}{\mathrm{P}(\hat{r}) + \mathrm{R}(\hat{r})}\ .$$

4.1.2 Rank-based retrieval

The set of all the possible ordered list of N retrieved documents is $D(N) = \{(d_1, \ldots, d_N) : d_i \in D, d_i \neq d_j \text{ for any } i \neq j\}$, i.e. a set of ranked lists of retrieved documents without duplicates. A run r is the vector $r = (d_1, \ldots, d_N)$ and we denote by $r[j]$ its jth element, i.e. $r[j] = d_j$. Similarly, a judged run is the vector $\hat{r} = \big(GT(t, d_1), \ldots, GT(t, d_N)\big)$, i.e. an ordered list of relevance degrees, where we denote by $\hat{r}[j]$ its jth element, i.e. $\hat{r}[j] = GT(t, d_j)$.

Let us introduce the definitions of some of the most popular rank-based measures:

- *Average Precision (AP)* (Buckley and Voorhees 2005) is a binary measure given by

$$\mathrm{AP}(\hat{r}) = \frac{1}{RB}\sum_{i=1}^{N}\left(\frac{1}{i}\sum_{j=1}^{i} g(\hat{r}[j])\right)g(\hat{r}_t[i]),$$

  where $g(\mathscr{a}_1) = 1$ and RB is the recall base.
- Defined $p \in (0, 1)$ the persistence parameter, *Graded Rank-Biased Precision (gRBP)* (Moffat and Zobel 2008; Sakai and Kando 2008) is a multi-graded relevance measure given by

$$\text{gRBP}(\hat{r}) = \frac{(1-p)}{g(\mathscr{a}_c)} \sum_{i=1}^{N} p^{i-1} g(\hat{r}[i]).$$

Typical values of p are 0.5 for a very impatient user, 0.8 for a relatively patient user, and 0.95 for a user very persistent in deeply scanning the result list. gRBP coincides with Rank-Biased Precision (RBP) when binary judgments ($c = 1$) are considered and $g(\mathscr{a}_1) = 1$.

- *Discounted Cumulated Gain (DCG)* (Järvelin and Kekäläinen 2002) is a multi-graded relevance measure given by

$$\text{DCG}_b(\hat{r}) = \sum_{i=1}^{N} \frac{g(\hat{r}[i])}{\max\{1, \log_b i\}},$$

where base b of the logarithm is typically equal to 2 for an impatient user and to 10 for a patient user.

- *Expected Reciprocal Rank (ERR)* (Chapelle et al. 2009) is a cascaded multi-graded relevance measure, given by

$$\text{ERR}(\hat{r}) = \sum_{i=1}^{N} \frac{1}{i} x_i \prod_{j=1}^{i-1} (1 - x_j),$$

with the convention that $\prod_{i=1}^{0} = 1$ and x_k represents the probability that a user leaves their search after considering the document at position k. Here we adopt the typical setting $x_k = (2^{g(\hat{r}[k])} - 1)/2^{g(\mathscr{a}_c)}$.

4.2 Set-based measures

Let us start by introducing an order relation $\preceq$ on the set of judged runs. Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$, and let k be the biggest relevance degree at which the two runs differ for the first time, i.e. $k = \max\{j \leq c : \left|\{i : \hat{r}_i = \mathscr{a}_j\}\right| \neq \left|\{i : \hat{s}_i = \mathscr{a}_j\}\right|\}$. We strictly order any pair of distinct system runs as follows

$$\hat{r} \prec \hat{s} \;\Leftrightarrow\; \left|\{i : \hat{r}_i = \mathscr{a}_k\}\right| < \left|\{i : \hat{s}_i = \mathscr{a}_k\}\right|. \tag{1}$$

$R(N)$ is a totally ordered set with respect to the ordering $\preceq$ defined by (1). As for any totally order set, $R(N)$ is a poset consisting of only one maximal chain (the whole set); therefore it is *graded of rank* $|R(N)| - 1$, where $\left|R(N)\right| = \binom{N+c}{N}$ since it consists of all the N combinations of $c + 1 = |REL|$ objects with repetition. Since $R(N)$ is graded of rank $|R(N)| - 1$, there exists a unique *rank function* $\rho(\hat{r}) : R(N) \longrightarrow \mathbb{N}$ such that $\rho(\hat{0}) = 0$ and $\rho(\hat{s}) = \rho(\hat{r}) + 1$ if $\hat{s}$ covers $\hat{r}$:

$$\rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_{\mathscr{a}}(\hat{r}_j) + N - j}{N - j + 1}, \tag{2}$$

where $\hat{r} = \{\hat{r}_1, \ldots, \hat{r}_N\} \in R(N)$ with $\hat{r}_i \preceq \hat{r}_{i+1}$ for any $i < N$.

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in R(N)$ 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 $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$. $(R(N), \preceq_d)$ is a difference structure. Thus the rank function is an interval scale and we are able to define a new measure that follows:

Definition 2 The *Set-based total order (SBTO)* measure on $(R(N), \preceq_d)$ is:

$$\text{SBTO}(\hat{r}) = \rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_{\mathscr{a}}(\hat{r}_j) + N - j}{N - j + 1}. \tag{3}$$

This measure satisfies the condition imposed by Theorem 1. Thus, SBTO is an interval scale on $(R(N), \preceq_d)$.

Let us explore more deeply how the SBTO measure works. The first relevance degree immediately above not relevant, i.e. $\mathscr{a}_1$, always gives a constant contribution, independently from how many $\mathscr{a}_1$ documents are retrieved, since:

$$\binom{\delta_{\mathscr{a}}(\mathscr{a}_1) + N - j}{N - j + 1} = \binom{1 + N - j}{N - j + 1} = 1 \, .$$

However, when we consider higher relevance degrees, i.e. $\mathscr{a}_k$ with $k > 1$, the binomial coefficient strictly depends and changes on the basis of how many of them are retrieved. Indeed, $\delta_{\mathscr{a}}(\mathscr{a}_k)$ is constant for all the documents with the same relevance degree $\mathscr{a}_k$, but the term $N - j$ decreases as the number of $\mathscr{a}_k$ retrieved documents increases due to N being constant and j increasing, i.e. the binomial coefficient is decreasing in the number of $\mathscr{a}_k$ retrieved documents. In other terms, each additional $\mathscr{a}_k$ retrieved document gives a contribution smaller than the previously retrieved ones by a discount factor j. This somehow recalls the idea that relevance is a dynamic notion which changes as far as more relevant documents are inspected, see e.g. (Mizzaro 1997). It can also be considered as a consequence of the submodularity principle studied by Chapelle et al. (2011). As a consequence, given $\hat{r}, \hat{s} \in R(N)$, a replacement in $\hat{r}$ may have a different effect than the same replacement in $\hat{s}$, if the relevance degree of the new document is greater than $\mathscr{a}_1$.

4.2.1 Binary relevance case

When $c = 1$, i.e. in the binary relevance case, the ordering (1) just orders judged runs by how many relevant documents they retrieve, i.e. by their total mass of relevance:

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \sum_{i=1}^{N} \delta_{\mathscr{a}}(\hat{r}_i) \leq \sum_{i=1}^{N} \delta_{\mathscr{a}}(\hat{s}_i) \, ,$$

since there is only one relevant relevance degree $\mathscr{a}_1$.

Therefore the rank function becomes

$$\rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathscr{a}}(\hat{r}_i) = \text{M}(\hat{r}) \, .$$

This follows easily from (3), using the fact that $\delta_{\mathscr{a}}(\hat{r}_i) \in \{0, 1\}$ for any $i \leq N$ when $c = 1$.

Let now g be the gain function, and let us consider *Precision*

$$\mathrm{P}(\hat{r}) = \frac{1}{N}\sum_{i=1}^{N}\frac{g(\hat{r}_i)}{g(\mathscr{a}_1)} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathscr{a}}(\hat{r}_i) = \frac{\mathrm{M}(\hat{r})}{N},$$

since $g(\mathscr{a}_0) = 0 = \delta_{\mathscr{a}}(\mathscr{a}_0)$ and $c = 1$. Thus Precision is an interval scale, as it is a linear positive transformation of M.

Similarly, *Recall*

$$\mathrm{R}(\hat{r}) = \frac{1}{RB}\sum_{i=1}^{N}\frac{g(\hat{r}_i)}{g(\mathscr{a}_1)} = \frac{1}{RB}\sum_{i=1}^{N}\delta_{\mathscr{a}}(\hat{r}_i) = \frac{\mathrm{M}(\hat{r})}{RB}$$

is an interval scale.

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

$$\mathrm{F}(\hat{r}) = 2\frac{\mathrm{P}(\hat{r})\cdot\mathrm{R}(\hat{r})}{\mathrm{P}(\hat{r})+\mathrm{R}(\hat{r})} = \frac{2}{N+RB}\sum_{i=1}^{N}\delta_{\mathscr{a}}(\hat{r}_i) = \frac{2\mathrm{M}(\hat{r})}{N+RB}$$

is an interval scale as well.

4.2.2 Multi-graded relevance case

Neither Generalized Precision nor Generalized Recall are a positive linear transformation of M defined in (3). Indeed, in these measures, the individual contribution of each retrieved document $\hat{r}_j$ is independent from the contribution of any other retrieved document $\hat{r}_k$. However, the previous discussion on the measure defined in (3) pointed out that, for each relevance degree $\mathscr{a}_k$ with $k > 1$, the individual contribution of an $\mathscr{a}_k$ retrieved document depends on how many $\mathscr{a}_k$ retrieved documents there are in the run. Therefore neither gP nor gR are an interval scale, since they cannot be a linear transformation of M.

Moreover they are not even an ordinal scale which, again, implies they cannot be an interval scale too. Indeed, a measure M' is an ordinal scale on $R(N)$ if, for every $\hat{r}, \hat{s} \in R(N)$, the following statement is true:

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

Let us consider $\hat{r} = \{\mathscr{a}_1, \ldots, \mathscr{a}_1\}$ and $\hat{s} = \{\mathscr{a}_2, \mathscr{a}_0, \ldots, \mathscr{a}_0\}$, two runs of length N. We have $\hat{r} \prec \hat{s}$. Moreover, since gR and gP are both proportional to $\mathrm{G}(\hat{v}) := \sum_{i=1}^{N} g(\hat{v}_i)/g(\mathscr{a}_c)$, for any $\hat{v} \in R(N)$, we can prove that $\mathrm{G}(\cdot)$ is not on an ordinal scale with respect to the order (1). Since $g(\mathscr{a}_0) = 0$, $\mathrm{G}(\hat{r}) = Ng(\mathscr{a}_1)/g(\mathscr{a}_c)$ while $\mathrm{G}(\hat{s}) = g(\mathscr{a}_2)/g(\mathscr{a}_c)$. From the fact that the gain function g is a positive strictly increasing function and it is defined independently from the length N of the runs, by choosing a N big enough we can have $\mathrm{G}(\hat{r}) > \mathrm{G}(\hat{s})$.

4.3 Rank-based measures

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 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 a total ordering among runs.

Let $\hat{r}, \hat{s} \in R(N)$ 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

$$\hat{r} \prec \hat{s} \;\Leftrightarrow\; \hat{r}[k] \prec \hat{s}[k]\,. \tag{4}$$

This ordering prefers a single relevant document ranked higher to any number of relevant documents, with the same relevance degree or higher, ranked just below it; more formally,

$$(\hat{u}[1], \ldots, \hat{u}[m], \mathscr{a}_0, \mathscr{a}_c, \ldots, \mathscr{a}_c), \prec (\hat{u}[1], \ldots, \hat{u}[m], \mathscr{a}_j, \mathscr{a}_0, \ldots, \mathscr{a}_0),$$

for any $1 \leq j \leq c$, for any length $N \in \mathbb{N}$ and any $m \in \{0, 1, \ldots, N-1\}$. This is why we call it *strong top-heaviness*.

$R(N)$ is totally ordered with respect to $\preceq$ and is *graded of rank* $(c+1)^N - 1$. Therefore, there is a unique rank function $\rho : R(N) \longrightarrow \{0, 1, \ldots, (c+1)^N - 1\}$ which is given by:

$$\rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathscr{a}}(\hat{r}[i])(c+1)^{N-i}, \tag{5}$$

where $\delta_{\mathscr{a}}$ is the indicator function.

Let us set $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r}) = (\delta_{\mathscr{a}}(\hat{r}[1]), \ldots, \delta_{\mathscr{a}}(\hat{r}[N]))$. If we look at $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r})$ as a string, the rank function is exactly the conversion in base 10 of the number in base $c+1$ identified by $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r})$ and the ordering among runs $\preceq$ corresponds to the ordering $\leq$ among numbers in base $c+1$.

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in R(N)$ 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 $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$. $(R(N), \preceq_d)$ is a difference structure. As done before in the set-based case, an interval scale measure on $(R(N), \preceq_d)$ is given by the rank function itself.

Definition 3 The *Rank-Based Total Order (RBTO)* measure on $(R(N), \preceq_d)$ is:

$$\mathrm{RBTO}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathscr{a}}(\hat{r}[i])(c+1)^{N-i} \tag{6}$$

This measure satisfies the condition imposed by Theorem 1. Thus, RBTO is an interval scale on $(R(N), \preceq_d)$.

Let $G = \min_{j \in \{1,\ldots,c\}}(g(\mathscr{a}_j) - g(\mathscr{a}_{j-1}))/g(\mathscr{a}_c) > 0$ be the normalized smallest gap between the gain of two consecutive relevance degrees.

gRBP_p with $p > G/(G+1)$ and other IR measures—namely AP, DCG, and ERR—are not even an ordinal scale on $R(N)$, as the following example shows. Let $\hat{r} = (\mathscr{a}_1, \mathscr{a}_0, \mathscr{a}_2, \mathscr{a}_0, \mathscr{a}_1)$ and $\hat{s} = (\mathscr{a}_1, \mathscr{a}_1, \mathscr{a}_0, \mathscr{a}_0, \mathscr{a}_0)$ be two runs on $R(5)$ with $c = 2$ and let us use the indicator function δ as gain function g. We have $\hat{r} \preceq \hat{s}$. Then $\mathrm{DCG}_2(\hat{r}) = 1 + 2/\log_2 3 + 1/\log_2 5 > 1 + 1 = \mathrm{DCG}_2(\hat{s})$; $\mathrm{ERR}(\hat{r}) = 1/4 + 3/16 + 3/320 > 1/4 + 3/32 = \mathrm{ERR}(\hat{s})$; finally, since $g(\mathscr{a}_2) = \delta_{\mathscr{a}}(\mathscr{a}_2) = 2$, $\quad 2\mathrm{gRBP}_p(\hat{r}) = (1-p)(1+2p^2+p^4) > (1-p)(1+p) = 2\mathrm{gRBP}(\hat{s})$ for $p \gtrsim 0.454$, and such an example can be found for any other values of $p > G/(G+1)$, where $G = 1/2$. AP is a binary measure and, just to stay with the same data above, we adopt a lenient mapping of multi-graded to binary relevance degrees setting $g(\mathscr{a}_1) = g(\mathscr{a}_2) = 1$ and thus $RB{\cdot}\mathrm{AP}(\hat{r}) = 1 + 2/3 + 3/5 > 1 + 1 = RB{\cdot}\mathrm{AP}(\hat{s})$, where RB is the recall base.

As a consequence, being not an ordinal scale, gRBP_p with $p > G/(G+1)$, AP, DCG, and ERR cannot be an interval scale too, since an interval scale measure is also an ordinal scale.

gRBP_p with $p \leq G/(G+1)$ is interval if and only if $p = G/(1+G) = (c+1)^{-1}$ and the gain function is equal to $g(\mathscr{a}_i) = K\delta_{\mathscr{a}}(\mathscr{a}_i)$, for any $i \in \{0, \ldots, \mathbb{N}\}$ and for any $K > 0$ fixed.

4.3.1 Summary of main findings and discussion

We summarize here the main findings of our framework:

- set-based evaluation measures:
  - binary relevance: precision, recall, F-measure are interval scales;
  - multi-graded relevance: gP and gR are neither ordinal nor interval scales;
- rank-based evaluation measures:
  - binary relevance: RBP is an interval scale only for $p = 1/2$ and it is an ordinal scale for $p < 1/2$; RBP for $p > 1/2$ and AP are neither ordinal nor interval scales;
  - multi-graded relevance: gRBP is an interval scale only for $p = G/(G+1)$ and when the gain function is equal to $g(\mathscr{a}_i) = K\delta_{\mathscr{a}}(\mathscr{a}_i)$; gRBP is an ordinal scale when $p < G/(G+1)$; gRBP for $p > G/(G+1)$, DCG, and ERR are neither ordinal nor interval scales.

Note that the relevance degrees are requested to be an ordinal scale and, being the gain function a monotone transformation of them, it is at least an ordinal scale. The above results ensure that measures are interval scales (or not) independently from additional properties of the gain function (provided it is at least an ordinal scale).

Carterette (2011) has shown how evaluation measures can be framed in terms of a browsing model, a document utility model (i.e. the gain function in our context), and a utility accumulation model. Moreover, he has shown how evaluation measures can be expressed as expectations of these utility models. Therefore, to reconcile our framework with the one by Carterette (2011) and to compute such expectations, it would be required that the gain function is an interval scale as well.

As a final remark, in the part of our framework (Ferrante et al. 2019) not reported in this paper, we formally identify conditions when the gain function must be a ratio scale (thus also an interval scale) in order to ensure that an evaluation measure can be on an interval scale. Therefore, these other cases can be used to determine when the gain function needs to be an interval scale in the Carterette (2011) sense.

5 Experiments

5.1 Experimental setup

We used the following *Text REtrieval Conference (TREC)* collections:

- Adhoc track `T08` (Voorhees and Harman 1999): 528,155 documents of the `TIPSTER` disks 4–5 corpus minus congressional record; `T08` provides 50 topics, each with binary relevance judgments and a pool depth of 100; 129 system runs were submitted to it;

- Core track T26 (Allan et al. 2018b): 1,855,658 documents of the New York Times from 1987 to 2007; T26 provides 50 topics, each with ternary relevance judgments and a pool depth up to 100, using both top-k and multi-armed bandits pooling techniques; 75 system runs were submitted to it.

T08 is used for binary relevance measures in both the set-based and rank-based cases while T26 is used for multi-graded relevance measures in both the set-based and rank-based cases. We have trimmed the length of the runs to 250 documents, since the definition of RBTO in Eq. (3) involves the power of the number of relevance degrees to the length of the run and this may cause overflow in some of the follow-up analyses when the length of the run is high. We have however validated this choice by comparing the measure scores with those of runs of length 100, 500, and 1000 and we found they are consistent with those of runs of length 250.

For SBTO and RBTO, in the multi-graded relevance case, we used the relevance weights $W_1 = [0, 1, 2]$ for not relevant, relevant, and highly relevant documents which correspond to the indicator function $\delta_{\mathscr{A}}(\mathscr{A}_i)$; in the case of RBTO we also experimented with two alternatives, one multiple of the indicator function $W_2 = [0, 2, 4] = 2W_1$ and the other not equi-spaced $W_3 = [0, 1, 3]$; note that SBTO in Eq. (2) depends only on the indicator function and so you cannot use alternative weighting schemes.

We used measures for both binary and multi-graded relevance measures:

- Binary relevance:
  - set-based measures: precision, recall, F-measure;
  - rank-based measures: *Average Precision (AP)* (Buckley and Voorhees 2005) and *Rank-Biased Precision (RBP)* (Moffat and Zobel 2008);
- Multi-graded relevance:
  - set-based measures: *Generalized Precision (gP)* and *Generalized Recall (gR)* (Kekäläinen and Järvelin 2002); we used the weights $W_1 = [0, 1, 2]$, which correspond to the indicator function.
  - rank-based measures: *Graded Rank-Biased Precision (gRBP)* (Moffat and Zobel 2008; Sakai and Kando 2008), *Discounted Cumulated Gain (DCG)* (Järvelin and Kekäläinen 2002), and *Expected Reciprocal Rank (ERR)* (Chapelle et al. 2009). For gRBP we used the weights $W_1 = [0, 1, 2]$, which correspond to the indicator function, but we experimented also with weights $W_3 = [0, 1, 3]$. For DCG and ERR we use their standard weights [0, 5, 10][1]; for DCG, we use a log_{10} discounting function, which accounts for a reasonably persistent user.

Note that when we do not indicate it explicitly, we intend that the weights $W_1 = [0, 1, 2]$ are used.

We considered a confidence level $\alpha = 0.05$ to determine if a factor is statistically significant.

We conducted three different types of analyses:

[1] We also experimented with the weights $W_1 = [0, 1, 2]$ to use exactly the same as those used in the case of RBTO and this produced very close experimental results, which are omitted for space reasons, preferring to use their standard weights for DCG and ERR.

- *correlation analysis*, reported in Sect. 5.2, is aimed at understanding the relationship among the different evaluation measures and how these are affected by their scales;
- *multiple comparison analysis*, reported in Sect. 5.3, is aimed at understanding the impact of violating or not the scale assumptions behind statistical tests for comparing IR systems;
- *incomplete information analysis*, reported in Sect. 5.4, is aimed at understanding how much less and less complete pools affect evaluation measures and to what extent the used scales play a role in this.

To ease the reproducibility of the experiments, the source code for running them is available at https://bitbucket.org/frrncl/irj2019-ffl/.

5.2 RQ1: What is the relationship between the different evaluation measures and how are these affected by their scales?

In order to address RQ1, we employ correlation analysis (Voorhees 1998), one of the most widely used tools to study properties and relationships among evaluation measures. The most used correlation coefficients are the Kendall's tau correlation τ (Kendall 1948) and the AP correlation τ_{AP} (Yilmaz et al. 2008). Ferro (2017) has shown that, when it comes to study evaluation measures, τ and τ_{AP} produce different absolute values yet ranking evaluation measures in the same way and, therefore, they lead to a consistent assessment. Thus, in the following, we report only Kendall's τ.

Given two rankings X and Y, their Kendall's τ correlation is given by

$$\tau(X, Y) = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}} \tag{7}$$

where P is the total number of concordant pairs (pairs that are ranked in the same order in both vectors), Q the total number of discordant pairs (pairs that are ranked in opposite order in the two vectors), T and U are the number of ties, respectively, in the first and in the second ranking.

$\tau \in [-1, 1]$ where $\tau = 1$ indicates two perfectly concordant rankings, i.e. in the same order, $\tau = -1$ indicates two fully discordant rankings, i.e. in opposite order, and $\tau = 0$ means that 50% of the pairs are concordant and 50% discordant.

The typical way of performing correlation analysis is as follows: let m_1 and m_2 be two evaluation measures; for example, let m_1 be AP. Let M_1 and M_2 be two $T \times S$ matrices where each cell contains the performances on topic i of system j according to measures m_1 and m_2, respectively. Therefore, M_1 and M_2 represent the performances of S different systems (columns) over T topics (rows); for example, M_1 contains the AP score of each system on each topic. Let $\overline{M}_1$ and $\overline{M}_2$ be the column-wise averages of the two matrices; for example, $\overline{M}_1$ is a vector where each element is the *Mean Average Precision (MAP)* of a system. If you sort systems by their score in $\overline{M}_1$ and $\overline{M}_2$, you obtain two *Rankings of Systems (RoS)* corresponding to m_1 and m_2, respectively. The Kendall's τ is then used to quantify how "close" these two RoS are. We call this approach *overall* since it first computes the average performance across the topics and then it computes the correlation among evaluation measures.

Note that the framework introduced in Sects. 4.2 and 4.3 holds topic-by-topic, e.g. two interval scale measures will order systems in the same way on the same topic and their

Table 1 Kendall's τ correlation analysis among set-based evaluation measures. The topic-by-topic score is the average across the topics

Binary relevance—T08		
Measure pair	Topic-by-topic	Overall
Precision versus SBTO	1.0000	0.9998
Recall versus SBTO	1.0000	0.8591
F-measure versus SBTO	1.0000	0.9670
Precision versus recall	1.0000	0.8588
SBTO versus RBTO	0.4358	0.7410
Multi-graded relevance—T26		
Measure pair	Topic-by-topic	Overall
Generalized precision versus SBTO	0.7325	0.9175
Generalized recall versus SBTO	0.7325	0.8453
Generalized precision versus generalized recall	1.0000	0.9003
SBTO versus RBTO	0.3895	0.7352

correlation will be 1.0. However, this may be no more true, if you first average performance across all the topics. Moreover, strictly speaking, measures which are not on an interval scale should be not averaged and this is exactly the first step in the computation of the overall correlation.

Therefore, we introduce a *topic-by-topic* way of computing correlation, which suits better with our framework. In this approach, for each topic i we consider the RoS on that topic corresponding to m_1 and m_2, i.e. we consider the ith rows of M_1 and M_2, respectively; we then compute the Kendall's τ correlation among the two RoS on that topic. Therefore, we end-up with a set of T correlation coefficients, one of each topic, which are then summarized considering their mean[2].

5.2.1 Set-based measures

Table 1 reports the correlation analysis in the case of set-based evaluation measures for both binary and multi-graded relevance.

As expected, in the binary case, the topic-by-topic correlation among precision, recall, F-measure, and SBTO is 1.00, since they are all on the same interval scale and they are just linear transformations one of the other. However, it is interesting to note how the overall correlation among precision and SBTO is 0.99 (it would be 1.00 but due to small approximations it is slightly different), while the one between recall and SBTO is 0.8591. This is due to the fact that, being precision and SBTO independent from the recall base, they basically order systems in a consistent way across the topics and this is reflected in the fact that

[2] Note that averaging Kendall's τ values implicitly assumes them to be on an interval scale and determining whether Kendall's τ is or not an interval scale goes beyond the scope of this paper. In the following, we consider the averaged Kendall's τ value more as a proxy to know whether all the topic-by-topic values are 1, i.e. whether we have an interval scale, or not.

the overall correlation is 1.0 in both cases. On the other hand, recall heavily depends on the recall base which changes for each topic and it is used to normalize the score for each topic; therefore, in a sense, recall on each topic changes the way it orders system and this is reflected in the overall correlation dropping to 0.85. F-measure, being the harmonic mean of precision and recall, falls somehow in-between and this effect is smoothed leading to an overall correlation of 0.96. Finally, we can note how the correlation among precision and recall behaves exactly in the same way as the correlation among them and SBTO, which is a sort of litmus test since precision and SBTO are substantially interchangeable.

These observations should also make us think about the way in which we typically interpret Kendall's τ overall correlation scores. The rule-of-thumb (Voorhees 1998, 2000) is that an overall correlation above 0.9 means that two evaluation measures are practically equivalent, an overall correlation between 0.8 and 0.9 means that two measures are similar, while dropping below 0.8 indicates that measures are departing more and more. However, these indications have been drawn analysing the problem of inter-assessor agreement and how much the same measure computed over the pools of different assessors agrees with itself. The analyses in Table 1 show that, in the case of precision, recall, and SBTO, we obtain an overall correlation of just 0.85 even if we know that they actually are on the same interval scale and thus we would have expected a higher overall correlation score, well above 0.9. Moreover, this marked difference in the overall correlation among them is not due to any considerable difference in the way they look at and order systems, but just to the way in which they normalize (or not) across topics. Leaving apart the debate on how evaluation measures heavily depending on the recall base are appropriate, whose value is at best a very rough estimation, these considerations may suggest that the topic-by-topic correlation analysis could be a good companion tool to adopt to study the behaviour of different evaluation measures, since the overall correlation may be affected by factors other than the user models behind evaluation measures and how they order systems.

More as a curiosity, the topic-by-topic correlation between SBTO and RBTO is 0.43, while the overall one is 0.74. This gives us a feeling of how big is the impact of passing from a set-based to a rank-based viewpoint. This difference between the set-based and rank-based viewpoints produces a fairly low topic-by-topic correlation, which is a bit higher in the case of the overall correlation, still being in the area of quite diverse measures.

In the case of multi-graded relevance, as expected, the topic-by-topic correlation among gP, gR and SBTO is low, since gP and gR are not an interval scale, and actually they are not even an ordinal scale. When it comes to the overall correlation, we can observe the same phenomenon due to the normalization (or not) by the recall base, since the overall correlation is higher between gP and SBTO than between gR and SBTO. In particular, the overall correlation between gP and SBTO is 0.91, which, according to the above rule-of-thumb, would lead us to consider the two measures practically equivalent. However, we know that SBTO and gP are substantially different and this underlines once more the issue on how we should interpret overall correlation scores.

On the other hand, we can observe as the topic-by-topic correlation between gP and gR is 1.00, indicating that they order systems in the same way and in accordance with the fact that they are just a linear transformation of one into the other. When it comes to the overall correlation, we can spot again the effect of the normalization (or not) by the recall base, since it drops to 0.90.

Finally, the topic-by-topic correlation between SBTO and RBTO, i.e. a proxy of the difference between the set-based and rank-based viewpoints, is 0.38, more than 12% lower than in the binary case. This is probably due to the additional complexity of the multi-graded case which substantially injects an additional type of ranking, i.e. the order among

Table 2 Correlation analysis among rank-based evaluation measures. The topic-by-topic score is the average across the topics

Binary relevance—T08		
Measure pair	Topic-by-topic	Overall
RBP $p = 1/2$ versus RBTO	1.0000	1.0000
RBP $p = 0.2$ versus RBTO	0.9985	0.9225
RBP $p = 0.8$ versus RBTO	0.8553	0.9043
AP versus RBTO	0.6099	0.7439
Multi-graded relevance—T26		
Measure pair	Topic-by-topic	Overall
gRBP $p = 1/3$ versus RBTO	1.0000	1.0000
gRBP $p = 1/3, W_3 = [0, 1, 3]$ versus RBTO	0.9867	0.9618
gRBP $p = 0.2$ versus RBTO	0.9996	0.9755
gRBP $p = 0.8$ versus RBTO	0.7420	0.9026
DCG versus RBTO	0.3774	0.6984
ERR versus RBTO	0.9468	0.9502
RBTO $W_1 = [0, 1, 2]$ versus RBTO $W_2 = [0, 2, 4]$	1.0000	1.0000
RBTO $W_1 = [0, 1, 2]$ versus RBTO $W_3 = [0, 1, 3]$	0.9866	0.9618

relevance degrees, which amplifies the differences. We can note how the overall correlation is less sensitive, changing just 0.8% with respect to the binary case and, again, this leads to the question whether it is the most appropriate tool for this kind of analyses.

5.2.2 Rank-based measures

Table 2 reports the correlation analysis in the case of rank-based evaluation measures for both binary and multi-graded relevance.

As expected, in the binary relevance case, the correlation between RBP with $p = \frac{1}{2}$ and RBTO is 1.00 since they are on the same interval scale and they are a linear transformation one of the other. Moreover, the correlation is 1.00 for both overall and topic-by-topic correlation because of the same line of reasoning on what kind of normalization is applied (or not) across topics, as discussed in the previous section. The topic-by-topic correlation between RBP with $p = 0.2$ and RBTO is 0.99, which is actually a small approximation for 1.00. Indeed, RBP with $p = 0.2$ is no more an interval scale, but it is still on ordinal scale and both RBP with $p = 0.2$ and RBTO keep ordering the systems in the same way, since RBTO is an ordinal scale too. The overall correlation drops to 0.92, being more affected by the difference in the scales, probably because in the case of the ordinal scale RBP with $p = 0.2$ you should not average the values, which is the preliminary step of the overall correlation. If we consider RBP with $p = 0.8$, which is not even an ordinal scale, the topic-by-topic correlation drops to 0.85. Overall, this shows how departing from an interval scale lowers more and more the correlation. In addition, in the case of RBP with $p = 0.8$ the overall correlation is 0.90, possibly suggesting a similarity between the measures bigger than what it actually is, since one is an interval scale while the other is not even an ordinal one; again, this raises the question on how to appropriately interpret overall correlation values.

Finally, the topic-by-topic correlation between AP and RBTO is 0.60 and this can be due to several factors: first, AP is not even an ordinal scale; then, AP normalizes scores across topics by the recall base; and, finally, the user models of AP and RBTO are different.

When it comes to multi-graded relevance, as expected, the correlation, both topic-by-topic and overall, between gRBP with $p = 1/3$ and RBTO is 1.00 since they are both interval scales and they are one the transformation of the other. However, as explained in Sect. 4.3 gRBP with $p = 1/3$ is an interval scale only when $g(a_i) = K\delta_a(a_i)$; therefore, we experimented with the alternative set of weights $W_3 = [0, 1, 3]$ which does not comply with this constraint. We can accordingly observe that the topic-by-topic correlation drops to 0.98 and the overall one to 0.96, which is a somehow moderate effect due to the small departure from this assumption. As it happened in the binary case, the topic-by-topic correlation with gRBP with $p = 0.2$, which is an ordinal scale, is 0.99, again a small approximation for 1.00, since gRBP with $p = 0.2$ and RBTO keep ordering systems in the same way. Finally, the topic-by-topic correlation with gRBP with $p = 0.8$, which is not even an ordinal scale, is 0.74, while the overall one is 0.90. Therefore, we observe a behaviour in the multi-graded case consistent with the one of the binary case and similar considerations hold.

The topic-by-topic correlation between DCG and RBTO is 0.37 and the overall one 0.69, which is quite low as well. Beyond the fact that DCG is not even an ordinal scale, this is probably due to a remarkable difference in the user models of DCG and RBTO.

The correlation between ERR and RBTO is quite high—topic-by-topic correlation is 0.94 and overall correlation is 0.95—despite the fact that ERR is not even an ordinal scale. This is probably due to the fact that both RBTO and ERR are quite top-heavy evaluation measures and this characteristic may prevail over the violation of the scales.

Finally, we considered two alternative sets of weights for RBTO and compared them to the weighting scheme of the indicator function. $W_2 = [0, 2, 4]$ is just a multiple of the indicator function and, as expected, the correlation between RBTO $W_1 = [0, 1, 2]$ and RBTO $W_2 = [0, 2, 4]$ is 1.00, since they are both the same interval scale, apart from a transformation by a multiplicative constant. On the other hand, the correlation between RBTO $W_1 = [0, 1, 2]$ and RBTO $W_3 = [0, 1, 3]$ slightly drops – topic-by-topic correlation is 0.98 and overall correlation is 0.96—and this is due to the fact that they are both interval scales, but now slightly different interval scales and no more just a transformation of the same scale.

5.3 RQ2: What is the impact of violating the scale assumptions behind statistical significance tests for comparing IR systems?

As previously explained, the type of scale determines the kind of operations you can perform with the obtained values: ordinal scales allow for computation of ranks and median, while interval scales allow also for mean and variance. One of the most common task in IR evaluation is to compare IR systems to understand which ones are significantly better.

We consider two types of statistical tests:

- the Kruskal–Wallis test (Kruskal and Wallis 1952), a non-parametric test that compares the medians of the groups of data to determine if the samples come from the same population by using the ranks of the data, rather than numeric values, to compute the test statistics. This type of analysis is thus suitable for both ordinal and interval scales;

Table 3 Tukey HSD test for set-based evaluation measures using the Kruskal–Wallis test and ANOVA. Each cell contains the number of significantly different pairs detected and, within parenthesis, the ratio with respect to the total number of system pairs

Binary relevance—T08, 8256 system pairs compared		
Measure pair	Significantly different pairs	
	Kruskal–Wallis test	ANOVA
Precision	1566 (18.97%)	2785 (33.73%)
Recall	1748 (21.17%)	3259 (39.47%)
F-measure	1721 (20.85%)	3081 (37.32%)
SBTO	1566 (18.97%)	2785 (33.73%)
Multi-graded relevance—T26, 2775 system pairs compared		
Measure pair	Significantly different pairs	
	Kruskal–Wallis test	ANOVA
Generalized precision	438 (15.78%)	1242 (44.76%)
Generalized recall	527 (18.99%)	1327 (47.82%)
SBTO	354 (12.76%)	938 (33.80%)

- the *ANalysis Of VAriance (ANOVA)* (Rutherford 2011), a parametric test which tests the hypothesis that all group means are equal. This type of analysis is thus suitable for interval scales only.

Being a parametric test, ANOVA is more powerful than the Kruskal–Wallis test and, generally speaking, it is able to spot more differences among the compared systems.

We consider how many significantly different system pairs these two tests are able to detect among all the possible pairs of systems under examination and we study how these figures change across evaluation measures and their scales. However, when performing multiple comparisons, the probability of committing a Type I error increases with the number of comparisons, i.e. it is more probable to detect significantly different pairs when they should not be detected (Fuhr 2017). Therefore, we keep this controlled by applying the Tukey *Honestly Significant Difference (HSD)* test (Hochberg and Tamhane 1987; Tukey 1949). Tukey's method is used in both Kruskal–Wallis test and ANOVA to create confidence intervals for all pairwise differences, while controlling the family error rate. For a deeper discussion of the assumptions behind ANOVA on other significance tests, the effect sizes, the power, and multiple comparisons, please refer to Carterette (2012).

5.3.1 Set-based measures

Table 3 reports the results of the Tukey HSD test and the number of significantly different pairs detected for both the Kruskal–Wallis test and ANOVA in the case of set-based evaluation measures.

In the case of binary relevance, all the set-based measures are on an interval scale and so they are suitable for being used with both the Kruskal–Wallis test and ANOVA. We can observe that, as expected, they all detect a comparable number of significantly different pairs and that this number increases when ANOVA is used, since it is a more powerful test

Table 4 Tukey HSD test for rank-based evaluation measures using the Kruskal–Wallis test and ANOVA. Each cell contains the number of significantly different pairs detected and, within parenthesis, the ratio with respect to the total number of system pairs

Binary relevance—T08, 8256 system pairs compared		
Measure pair	Significantly different pairs	
	Kruskal–Wallis test	ANOVA
RBP $p = 1/2$	1677 (20.31%)	2861 (34.65%)
RBP $p = 0.2$	1675 (20.29%)	2198 (26.62%)
RBP $p = 0.8$	1783 (21.60%)	3476 (42.10%)
AP	1824 (22.09%)	3320 (40.21%)
RBTO	1677 (20.31%)	2861 (34.65%)
Multi-graded relevance—T26, 2775 system pairs compared		
Measure pair	Significantly different pairs	
	Kruskal–Wallis test	ANOVA
gRBP $p = 1/3$	254 (9.15%)	551 (19.86%)
gRBP $p = 0.2$	254 (9.15%)	471 (16.97%)
gRBP $p = 0.8$	301 (10.85%)	885 (31.89%)
DCG	426 (15.35%)	1,274 (45.91%)
ERR	248 (8.94%)	566 (20.40%)
RBTO	254 (9.15%)	551 (19.86%)

than Kruskal–Wallis. We can also note that recall and F-measure detect a slightly higher number of different pairs and this is probably due to the use of the recall base for normalizing across topics.

In the case of multi-graded relevance, where gP and gR are neither ordinal nor interval scales, we can observe how they detect a higher number of significantly different pairs than RBTO. While there might be also other motivations such as the power of the tests (Carterette 2012) or the discriminative power of the measures (Sakai 2006), we can consider this as a tendency also due to an overestimation of the number of significantly different pairs, since both gP and gR violate the scale assumptions behind both the Kruskal–Wallis test and ANOVA.

5.3.2 Rank-based measures

Table 4 reports the results of the Tukey HSD test and the number of significantly different pairs detected for both the Kruskal–Wallis test and ANOVA in the case of rank-based evaluation measures.

In the binary relevance case, RBTO and RBP with $p = 1/2$ are interval scales and they match the scale assumptions behind both the Kruskal–Wallis test and ANOVA. RBP with $p = 0.2$ is an ordinal scale and, therefore, it matches the scale assumptions for the Kruskal–Wallis test, but not for ANOVA. We can note how, in the case of the Kruskal–Wallis test, it detects more or less the same number of significantly different pairs while for ANOVA, being provided with a less powerful scale than the one assumed, it detects less

significantly different pairs. As before, on top of other factors as the power of the test and the discriminative power of an evaluation measure, this might suggest that violating the scale assumptions somehow leads to an underestimation of the number of significantly different pairs.

When it comes to RBP with $p = 0.8$ and AP, they are neither ordinal nor interval scales and we can observe a phenomenon we have seen also in the case of the set-based evaluation measures: they detect a higher number of significantly different pairs and, under the previous caveats, we may consider this as a sort of overestimation.

The multi-graded relevance case behaves in a consistent way as well. gRBP with $p = 0.2$ is an ordinal scale and, using the Kruskal–Wallis test, it detects the same number of significantly different pairs as gRBP with $p = 1/3$ and RBTO, which are interval scales. On the other hand, it detects less significantly different pairs when using ANOVA, something which we may consider as an underestimation, with the limitations discussed above, due to the fact that it violates the ANOVA scale assumptions and it relies on a less powerful scale.

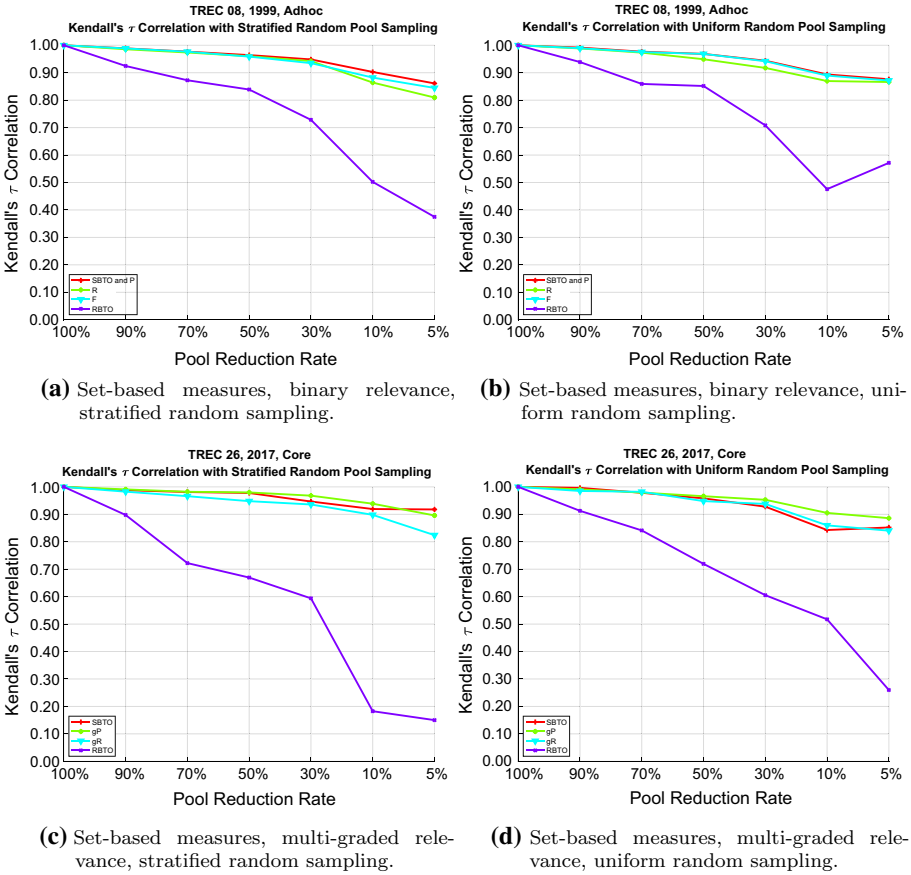
Finally, gRBP with $p = 0.8$, DCG, and ERR are neither ordinal nor interval scales and, as it happened before, they tend to detect a higher number of significantly different pairs, something we may consider as an overestimation, again considering the above caveats.

5.4 RQ3: How much do less and less complete pools affect evaluation measures and to what extent do the used scales play a role in this?

The downsampling pools allow us to investigate the behavior of evaluation measures as relevance judgments become less and less complete. We explore two pool sampling approaches:

- stratified random sampling (Buckley and Voorhees 2004): for each topic, a separate list of documents at each relevance grade (not relevant, relevant, highly relevant) is created from the original pool; for each sampling ratio $P\%$, we select $X = P\% \times D$ documents at the given relevance level, ensuring that at least 1 somehow relevant document and at least 10 not relevant documents are selected; the first $\max(1, X)$ documents from the random list at each relevant level have then been selected to constitute the new reduced pool; each smaller pool is a subset of each larger pool since we always select from the top of the lists. We used $P\% = [90, 70, 50, 30, 10, 5]$.
- uniform random sampling (Yilmaz and Aslam 2006): for each sampling ratio $P\%$, we uniformly select at random $X = P\% \times D$ documents from the pool, regardless of their relevance degree; if the random sample does not contain any relevant document, it is thrown away and another one is drawn. Also in this case, we used $P\% = [90, 70, 50, 30, 10, 5]$.

The plots in the following figures show the Kendall's τ correlations between the RoS produced using progressively down-sampled pools from 100% (complete pool) to 5%. Each line shows the behavior of a measure; the flatter (and closer to 1.0) the line, the more a measure ranks systems in the same relative order with different levels of relevance judgments incompleteness.



(a) Set-based measures, binary relevance, stratified random sampling.

(b) Set-based measures, binary relevance, uniform random sampling.

(c) Set-based measures, multi-graded relevance, stratified random sampling.

(d) Set-based measures, multi-graded relevance, uniform random sampling.

Fig. 1 Self Kendall's τ correlation at pool samples for set-based measures

5.4.1 Set-based measures

Figure 1 show the self Kendall's τ correlation at the different pool samples for the set-based evaluation measures for both binary and multi-graded relevance; on the left, there is the stratified random sampling and on the right there is the uniform random sampling.

In the case of binary relevance, we can observe as SBTO and precision behave in the same way, while recall has a slightly lower self-correlation; this is probably due to the estimation of the recall base which gets worse and worse as the sample size is reduced. F-measure performs in-between precision and recall. These trends are consistent among the two pool sampling strategies.

In the case of multi-graded relevance, SBTO, gP and gR behave in a very close way, even if for different reasons, since gP and gR are not on an interval scale. As in the binary case, gR has a slightly lower self-correlation and this is probably due to the estimate of the recall base, which becomes less and less accurate as the sample size is reduced.

Figure 1 also shows the self-correlation of RBTO, so that it is possible to get a feeling of what is the impact of passing from a set-based to a rank-based viewpoint, still remaining on an interval scale. We can see how the self-correlation of RBTO is substantially lower than

the one of SBTO, even more in the case of stratified random sampling. This behaviour is consistent in both the binary and multi-graded cases.

When you downsample the pool, you are basically reducing the number of relevant documents while keeping the length of the run the same; as a consequence, you are reducing the number of relevant documents that can appear in a run, increasing the number of not relevant ones. The total order behind SBTO and RBTO basically orders in an equi-spaced way the set of all the possible runs with a given number of relevant documents; when you reduce the number of relevant documents you also decrease the number of all the possible runs of a given length and this decrease is much more pronounced in the case of rank-based than set-based evaluation measures since, with the same number of relevant documents, the rank-based case originates a much bigger number of possible cases.

As a consequence, the same set of real runs submitted to a track is mapped to a space of possible runs which gets smaller and smaller as the sample size is reduced and this decrease is much sharper in the case of rank-based retrieval. Therefore, the same set of real runs is "conflated" to smaller spaces of possible runs and this may, for example, originate more ties and undistinguishable runs. Thus, this prevents, more and more, an interval scale measure to rank systems in the same way as on the full pool. Since this phenomenon is much more pronounced in the case of rank-based evaluation measures than of set-based ones, it happens out that the self-correlation decreases more for RBTO than for SBTO.

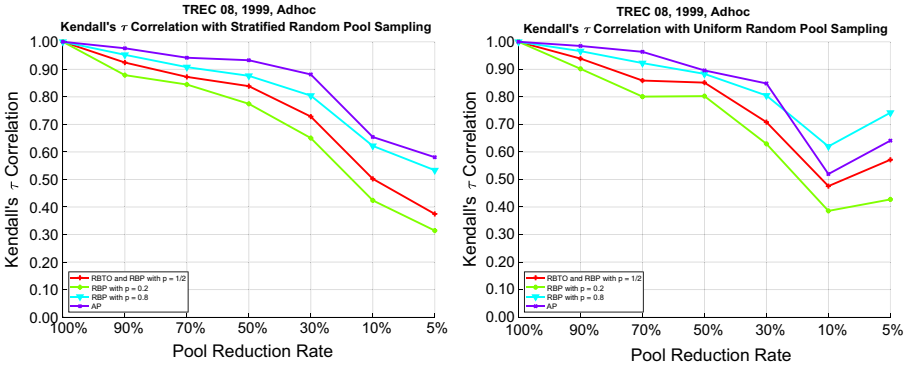
5.4.2 Rank-based measures

Figure 2 shows the self Kendall's τ correlation at the different pool samples for the rank-based evaluation measures for both binary and multi-graded relevance; on the left, there is the stratified random sampling and on the right there is the uniform random sampling.

As expected, RBTO and RBP with $p = 1/2$ in the binary case and gRBP with $p = 1/3$ in the multi-graded case behave in the same way, since they are on the same interval scale. We can also observe as both RBP with $p = 0.2$ and gRBP with $p = 0.2$ have a slightly lower self-correlation than RBTO and this can be explained by them being on an ordinal scale rather than an interval one. On the other hand, RBP with $p = 0.8$ and gRBP with $p = 0.8$, which are both neither interval nor ordinal scales, have a higher self-correlation than RBTO. This may be due to the "conflation" mechanism described above, which is less marked for RBP and gRBP with $p = 0.8$. This phenomenon is even more evident in the case of AP and DCG, which exhibit even higher self-correlations; it is a little bit less pronounced in the case of ERR since its strong top-heaviness makes it more sensible to a reduction in the pool. What do AP, DCG, and ERR see as actual space of the possible runs and how this space conflates or not, as pools are reduced, remain open questions whose answers may provide us with some additional insights on why they have higher self-correlation scores.

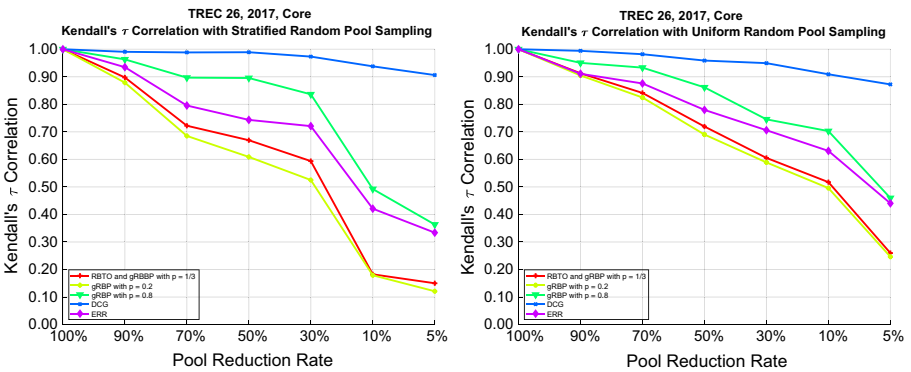
6 Conclusions and future work

In this paper, we stepped from our theory of IR evaluation measures and its definition of measurement scales to conduct an experimental study, based on standard TREC collections, aimed at assessing the impact of our theoretical findings with respect to state-of-the-art evaluation measures and some of the most common types of conducted analyses. Indeed, our formal framework allowed us to determine whether and when set-based and



(a) Rank-based measures, binary relevance, stratified random sampling.

(b) Rank-based measures, binary relevance, uniform random sampling.



(c) Rank-based measures, multi-graded relevance, stratified random sampling.

(d) Rank-based measures, multi-graded relevance, uniform random sampling.

Fig. 2 Self Kendall's τ correlation at pool samples for rank-based measures

rank-based IR measures are interval scales and this is a fundamental question since the validity of the descriptive statistics, such as mean and variance, and the statistical significance tests we daily use to compare IR systems depend on its answer.

We addressed RQ1 by conducting a correlation analysis to understand the relationship among evaluation measures and their scales. We found out that, as expected, when evaluation measures are on the same interval scale, their correlation is 1.00; this holds also in the case of the relationship between measures on interval and ordinal scales, whose correlation is still 1.00 because they keep ordering the systems in the same way. We have also shown how much the correlation drops when you compare measures which are on an interval scale to measures which are neither ordinal nor interval scales: this drop is not only due to differences in the user models embedded in the evaluation measures, but also due to the violation of the scale assumptions.

On a methodological side, we noted how the usual way of computing the correlation among evaluation measures, that we called overall correlation, may not be the most suitable one for studying scale properties, since its preliminary averaging operation may introduce biases, especially when used with measures which are neither ordinal nor

interval scales. Therefore, we introduced a topic-by-topic correlation analysis to more appropriately study scale properties.

We addressed RQ2 by performing a multiple comparison test analysis, which is typically used to compare IR systems and detect which are significantly different. We considered the Kruskal–Wallis test, which is a non parametric test comparing medians and suitable for ordinal (and interval) scales, and ANOVA, which is a parametric test comparing means and suitable for interval scales only. We found that, as expected, both ordinal and interval scale measures behave in a similar way when using the Kruskal–Wallis test, for which both of them are appropriate. On the other hand, when you violate the scale assumptions behind statistical significance tests, provided that other factors, such as the power of the test and the discriminative power of the evaluation measures, may play an important role, you can observe variations in the number of detected significantly different pairs, which may be due also to the lack of compliance with the scale assumptions. In particular, when you perform ANOVA using ordinal scale measures, they tend to somehow underestimate the number of significantly different pairs, since ordinal scales are less powerful than interval ones expected for ANOVA. Finally, when you use measures which are neither ordinal nor interval scales, they tend to overestimate the number of significantly different pairs, in the case of both the Kruskal–Wallis test and ANOVA.

Finally, we addressed RQ3 by performing an analysis with respect to incomplete information, i.e. when you downsample pools. We found that measures on the same interval scale behave in a similar way and that measures on ordinal scales tend to be more sensitive to incomplete information. Moreover, incomplete information impacts more rank-based than set-based measures on interval scales because the former ones suffer from a sharpest "conflation" in the space of the possible runs to be totally ordered. This may also be an explanation why rank-based evaluation measures, which are neither ordinal nor interval scales, are much less sensitive to pool downsampling than interval scale measures.

In this paper, we actually used only a part of our theory of IR evaluation measures, namely the one based on a total order among system runs. Indeed, this total order guarantees to work for any possible set of real runs, as the `T08` and `T26` runs are, independently from how sparse this sample of real runs is with respect to the set of all the possible runs of a given length. On the other hand, our theory contains also interval scales which are developed starting from a partial ordering among system runs. This means that only a subset of runs can be ordered together, i.e. we are working with posets, and that ordered runs in a poset are not comparable to ordered runs in another poset. This is challenging from an experimental point of view because a set of real runs is a very small sample of all the possible runs of a given length and you may end up having runs that belong to many different posets, at the extreme one run per poset, and these runs would not be directly comparable. Therefore, it would turn out to be practically very difficult to conduct an analysis similar to the one we did in this paper. As future work, we will thus investigate how actual runs are distributed across posets, trying to find out a viable way of analysing them; an option could be also to use a mix between real and simulated runs to avoid having too sparse data.

Finally, even if our work is focused on offline evaluation measures, the concordance of offline measures with user feelings (satisfaction, preference, etc.) and with online measures (e.g., number of clicks in a session, page click-through rate, number of clicks divided by the position of the lowest click, mean reciprocal ranks of the clicks) is a very relevant research area. Therefore, our future work will also consider the possibility of extending our framework to online measures as well as studying how our interval-based offline measures relate to online ones, using for example the approach adopted by Chen et al. (2017).

References

Allan, J., Arguello, J., Azzopardi, L., Bailey, P., Baldwin, T., Balog, K., et al. (2018a). Research frontiers in information retrieval—Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018). *SIGIR Forum*, *52*(1), 34–90.

Allan, J., Harman, D. K., Kanoulas, E., Li, D., Van Gysel, C., & Voorhees, E. M. (2018b). TREC 2017 common core track overview. In E. M. Voorhees, & A. Ellis (Eds.), *The twenty-sixth text retrieval conference proceedings (TREC 2017)*. National Institute of Standards and Technology (NIST), Special Publication 500-324, Washington, USA.

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, M. F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*(4), 461–486.

Amigó, E., Gonzalo, J., & Verdejo, M. F. (2013). A general evaluation measure for document organization tasks. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, & T. Sakai (Eds.), *Proceedings of 36th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2013)* (pp. 643–652). New York, USA: ACM Press.

Amigó, E., Fang, H., Mizzaro, S., & Zhai, C. (2017). Report on the SIGIR 2017 workshop on axiomatic thinking for information retrieval and related tasks (ATIR). *SIGIR Forum*, *51*(3), 99–106.

Amigó, E., Fang, H., Mizzaro, S., & Zhai, C. (2018a). Are we on the right track? An examination of information retrieval methodologies. In K. Collins-Thompson, Q. Mei, B. Davison, Y. Liu, & E. Yilmaz (Eds.), *Proceedings of 41th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2018)* (pp. 997–1000). New York, USA: ACM Press.

Amigó, E., & Spina, D. (2018b). An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In K. Collins-Thompson, Q. Mei, B. Davison, Y. Liu, & E. Yilmaz (Eds.), *Proceedings of 41th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2018)* (pp. 625–634). New York, USA: ACM Press.

Bollmann, P. (1984). Two axioms for evaluation measures in information retrieval. In C. J. van Rijsbergen (Ed.), *Proceedings of the third joint BCS and ACM symposium on research and development in information retrieval* (pp. 233–245). Cambridge, UK: Cambridge University Press.

Bollmann, P., & Cherniavsky, V. S. (1980). Measurement-theoretical investigation of the MZ-metric. In C. J. van Rijsbergen (Ed.), *Proceedings of 3rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1980)* (pp. 256–267). New York, USA: ACM Press.

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), *Proceedings of 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004)* (pp. 25–32). New York, USA: ACM Press.

Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In D. K. Harman, & E. M. Voorhees (Eds.), *TREC. Experiment and evaluation in information retrieval* (pp. 53–78). Cambridge, MA: MIT Press.

Busin, L., & Mizzaro, S. (2013). Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In O. Kurland, D. Metzler, C. Lioma, B. Larsen, & P. Ingwersen (Eds.), *Proceedings of 4th international conference on the theory of information retrieval (ICTIR 2013)* (pp. 22–29). New York, USA: ACM Press.

Carterette, B. A. (2011). System effectiveness, user models, and user utility: A conceptual framework for investigation. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, & W. B. Croft (Eds.), *Proceedings of 34th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2011)* (pp. 903–912). New York, USA: ACM Press.

Carterette, B. A. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 4:1–4:34.

Chapelle, O., Metzler, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, & J. J. Lin (Eds.), *Proceedings of 18th international conference on information and knowledge management (CIKM 2009)* (pp. 621–630). New York, USA: ACM Press.

Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L., & Wu, S.-L. (2011). Intent-based diversification of web search results: Metrics and algorithms. *Information Processing & Management*, *14*(6), 572–592.

Chen, Y., Zhou, K., Liu, Y., Zhang, M., & Ma, S. (2017). Meta-evaluation of online and offline web search evaluation metrics. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, & R. W. White (Eds.), *Proceedings of 40th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2017)* (pp. 15–24). New York, USA: ACM Press.

Ferrante, M., Ferro, N., & Maistro, M. (2015). Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, & Y. Zhang

(Eds.), *Proceedings of 1st ACM SIGIR international conference on the theory of information retrieval (ICTIR 2015)* (pp. 21–30). New York, USA: ACM Press.

Ferrante, M., Ferro, N., & Pontarollo, S. (2017a). Are IR evaluation measures on an interval scale? In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, & E. Yilmaz (Eds.), *Proceedings of 3rd ACM SIGIR international conference on the theory of information retrieval (ICTIR 2017)* (pp. 67–74). New York, USA: ACM Press.

Ferrante, M., Ferro, N., & Pontarollo S (2017b). An interval-like scale property for IR evaluation measures. In N. Ferro, & I. Soboroff (Eds.), *Proceedings of 8th international workshop on evaluating information access (EVIA 2017)* (pp. 10–15). CEUR Workshop Proceedings (CEUR-WS.org). ISSN 1613-0073, http://ceur-ws.org/Vol-2008/. Accessed June 2019.

Ferrante, M., Ferro, N., & Pontarollo, S. (2019). A general theory of IR evaluation measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *31*(3), 409–422.

Ferro, N. (2017). What does affect the correlation among evaluation measures? *ACM Transactions on Information Systems (TOIS)*, *36*(2), 19:1–19:40.

Ferro, N., Fuhr, N., Grefenstette, G., Konstan, J. A., Castells, P., Daly, E. M., et al. (2018). The Dagstuhl perspectives workshop on performance modeling and prediction. *SIGIR Forum*, *52*(1), 91–101.

Fuhr, N. (2017). Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, *51*(3), 32–41.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, USA: Wiley.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422–446.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology (JASIST)*, *53*(13), 1120–1129.

Kendall, M. G. (1948). *Rank correlation methods*. Oxford, England: Griffin.

Knuth, D. E. (1981). *The art of computer programming—Volume 2: Seminumerical algorithms* (2nd ed.). Reading, USA: Addison-Wesley.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. Additive and polynomial representations* (Vol. 1). New York, USA: Academic Press.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Maddalena, E., & Mizzaro, S. (2014). Axiometrics: Axioms of information retrieval effectiveness metrics. In S. Mizzaro & R. Song (Eds.), *Proceedings of 6th international workshop on evaluating information access (EVIA 2014)* (pp. 17–24). Tokyo, Japan: National Institute of Informatics.

Miyamoto, S. (2004). Generalizations of multisets and rough approximations. *International Journal of Intelligent Systems*, *19*(7), 639–652.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science and Technology (JASIST)*, *48*(9), 810–832.

Moffat, A. (2013). Seven Numeric Properties of Effectiveness Metrics. In R. E. Banchs, F. Silvestri, T.-Y. Liu, M. Zhang, S. Gao, & J. Lang (Eds.), *Proceedings of 9th Asia information retrieval societies conference (AIRS 2013)*. Lecture Notes in Computer Science (LNCS) 8281 (Vol. 8281, pp. 1–12). Springer, Heidelberg

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, *27*(1), 2:1–2:27.

Rossi, G. B. (2014). *Measurement and probability. A probabilistic theory of measurement with applications*. New York, USA: Springer.

Rutherford, A. (2011). *ANOVA and ANCOVA. A GLM approach* (2nd ed.). New York, USA: Wiley.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. In E. N. Efthimiadis, S. Dumais, D. Hawking, & K. Järvelin (Eds.), *Proceedings of 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2006)* (pp. 525–532). New York, USA: ACM Press.

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, *11*(5), 447–470.

Stanley, R. P. (2012). *Enumerative combinatorics—Volume 1, volume 49 of Cambridge Studies in Advanced Mathematics* (2nd ed.). Cambridge, UK: Cambridge University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, New Series*, *103*(2684), 677–680.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99–114.

van Rijsbergen, C. J. (1974). Foundations of evaluation. *Journal of Documentation*, *30*(4), 365–373.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London, England: Butterworths.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1998)* (pp. 315–323). New York, USA: ACM Press.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, *36*(5), 697–716.

Voorhees, E. M., & Harman, D. K. (1999). Overview of the eigth text retrieval conference (TREC-8). In E. M. Voorhees, & D. K. Harman (Eds.), *The eighth text retrieval conference (TREC-8)* (pp. 1–24). National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA.

Yilmaz, E., & Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments. In P. S. Yu, V. Tsotras, E. A. Fox, & C.-B. Liu (Eds.), *Proceedings of 15th international conference on information and knowledge management (CIKM 2006)* (pp. 102–111). New York, USA: ACM Press.

Yilmaz, E., Aslam, J. A., & Robertson, S. E. (2008). A new rank correlation coefficient for information retrieval. In T.-S. Chua, M.-K. Leong, D. W. Oard, & F. Sebastiani (Eds.), *Proceedings of 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2008)* (pp. 587–594). New York, USA: ACM Press.