A formal theory to determine scale properties of evaluation measures*

Marco Ferrante[1], Nicola Ferro[2], and Silvia Pontarollo[1]

[1] Department of Mathematics, University of Padua, Italy
{ferrante, spontaro}@math.unipd.it
[2] Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it

Abstract. Evaluation measures are the basis for quantifying the performance of information access 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.
We define a formal theory of evaluation measures, based on the representational theory of measurement, which allows us to determine whether and when 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 measures are not interval scales.

1 Introduction

Measurement scales play a central role [4,5] since they determine the operations that can be performed with the measured values and, as a consequence, the statistical analyses that can be applied. Stevens [5] identifies four major types of scales with increasing properties: (i) the *nominal scale* consists of discrete unordered values, i.e. categories; (ii) the *ordinal scale* introduces a natural order among the values; (iii) the *interval scale* preserves the equality of intervals or differences; and (iv) the *ratio scale* preserves the equality of ratios.

In experimental evaluation we daily perform operations, such as computing means and variances, which are also the basic "ingredients" of the more sophisticated statistical significance tests we use to compare systems and assess their differences [1]. However, all these operations can be performed only from interval

* Extended abstract of [2]

scales onwards but, due to our limited knowledge of evaluation measures, we do not actually know which scales they rely on.

This paper sets a theory of evaluation measures to formally investigate their properties and to study whether and when they use an interval scale. We frame our work within the *representational theory of measurement* [4], which is the measurement theory adopted in both physical and social sciences. In particular, we develop a fully comprehensive framework, comprising both *set-based* measures – namely Precision, Recall and F-measure – dealing with unordered result lists, and *rank-based* measures – namely, *Average Precision (AP)*, *Discounted Cumulated Gain (DCG)*, *Rank-Biased Precision (RBP)*, and *Expected Reciprocal Rank (ERR)* – dealing with ranked result lists. Moreover, we consider both *binary relevance*, i.e. when documents can be just relevant or not relevant, and *multi-graded relevance* [3], i.e. when documents can have different degrees of relevance, such as not relevant, partially relevant, and highly relevant.

The paper is organized as follows. Section 2 introduces how to determine when a measure is an interval scale, according to the representational theory of measurement; Section 3 introduces our basic formalisms; Sections 4 and 5 discuss set-based and rank-based evaluation measures, respectively; finally, Section 6 wraps up the discussion and outlooks some future work.

2 Representational Theory of Measurement

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 $\preceq$. 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".

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

i. $\preceq_d$ *is* weak order*;*
ii. *if* $\Delta_{ab} \preceq_d \Delta_{cd}$*, then* $\Delta_{dc} \preceq_d \Delta_{ba}$;

iii. Weak Monotonicity: *if* $\Delta_{ab} \preceq_d \Delta_{a'b'}$ *and* $\Delta_{bc} \preceq_d \Delta_{b'c'}$ *then* $\Delta_{ac} \preceq_d \Delta_{a'c'}$;
iv. Solvability Condition*: if* $\Delta_{aa} \preceq_d \Delta_{cd} \preceq_d \Delta_{ab}$*, then there exists* $d', d'' \in R$ *such that* $\Delta_{ad'} \sim_d \Delta_{cd} \sim_d \Delta_{d''b}$*.*

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

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 evaluation measures are interval scales or not, [2] 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 evaluation measures are a linear positive transformation of this interval scale M and determine whether they are an interval scale.

3 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 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, $g(a_0) = 0$ and we require g to be strictly increasing.

We define the **indicator function** for the relevance degrees as $\delta_a(a_j) = j \ \ \forall j \in \{0, \dots, c\}$. Note that δ_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 $REL = \{a_0, a_1\}$, the gain function usually is $g(a_1) = \delta_a(a_1) = 1$ 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.

4 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. $k = \max\{j \leq c : \big|\{i : \hat{r}_i = a_j\}\big| \neq \big|\{i : \hat{s}_i = a_j\}\big|\}$. We strictly order any pair of distinct system runs as follows

$$\hat{r} \prec \hat{s} \ \Leftrightarrow \ \big|\{i : \hat{r}_i = a_k\}\big| < \big|\{i : \hat{s}_i = a_k\}\big| \ . \tag{1}$$

$R(N)$ 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 $\big|R(N)\big| = \binom{N+c}{N}$ 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 $\hat{r}$:

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

where $\hat{r} = \{\hat{r}_1, \dots, \hat{r}_N\} \in R(N)$ 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:*

$$\mathrm{SBTO}(\hat{r}) = \rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_a(\hat{r}_j) + N - j}{N - j + 1} . \tag{3}$$

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. a_1, always gives a constant contribution, independently from how many a_1 documents are retrieved, since:

$$\binom{\delta_a(a_1) + N - j}{N - j + 1} = \binom{1 + N - j}{N - j + 1} = 1 .$$

However, when we consider higher relevance degrees, i.e. a_k with $k > 1$, the binomial coefficient strictly depends and changes on the basis of how many of them are retrieved. Indeed, $\delta_a(a_k)$ is constant for all the documents with the same relevance degree a_k, but the term $N - j$ decreases as the number of a_k retrieved documents increases due to N being constant and j increasing, i.e. the binomial coefficient is decreasing in the number of a_k retrieved documents. In other terms, each additional a_k retrieved document gives a contribution smaller than the previously retrieved ones by a discount factor j.

4.1 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:

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

since there is only one relevant relevance degree a_1.

Therefore the rank function becomes

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

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

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

$$\textit{Precision } (P)(\hat{r}) = \frac{1}{N} \sum_{i=1}^{N} \frac{g(\hat{r}_i)}{g(a_1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_a(\hat{r}_i) = \frac{\mathrm{M}(\hat{r})}{N} ,$$

since $g(a_0) = 0 = \delta_a(a_0)$ and $c = 1$. Thus Precision is an interval scale, as it is a linear positive transformation of M.

Similarly, *Recall*

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

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 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 *Generalized Recall (gR)* and *Generalized Precision (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 $\mathrm{G}(\hat{r}) > \mathrm{G}(\hat{s})$.

5 Rank-based Measures

Top-heaviness is a central property in *Information Retrieval (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 order prefers a single relevant document ranked higher to any number of relevant documents, with same relevance degree or higher, ranked just below it

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

for any $1 \leq j \leq c$, for any length $N \in \mathbb{N}$ and any $m \in \{0, 1, \dots, N-1\}$. 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 $\rho : R(N) \longrightarrow \{0, 1, \dots, (c+1)^N - 1\}$ 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 $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r}) = (\delta_{\mathscr{a}}(\hat{r}[1]), \dots, \delta_{\mathscr{a}}(\hat{r}[N]))$. 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 $G = \min_{j \in \{1,\dots,c\}} (g(\mathscr{a}_j) - g(\mathscr{a}_{j-1}))/g(\mathscr{a}_c) > 0$ be the normalized smallest gap between the gain of two consecutive relevance degrees.

Graded Rank-Biased Precision (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 $g(\mathscr{a}_2) = \delta_{\mathscr{a}}(\mathscr{a}_2) = 2$, $2\mathrm{gRBP}_p(\hat{r}) = (1-p)(1 + 2p^2 + p^4) > (1-p)(1+p) = 2\mathrm{gRBP}(\hat{s})$ 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 $g(\mathpzc{a}_1) = g(\mathpzc{a}_2) = 1$ 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 $g(\mathpzc{a}_i) = K\delta_{\mathpzc{a}}(\mathpzc{a}_i)$, for any $i \in \{0, \dots, \mathbb{N}\}$ and for any $K > 0$ fixed.

6 Conclusion and Future Works

We developed a theory of evaluation measures to explore whether and when both set-based and rank-based IR measures are interval scales. 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 depends on its answer.

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 $g(\mathpzc{a}_i) = K\delta_{\mathpzc{a}}(\mathpzc{a}_i)$; 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.

As future work, we will investigate these new interval scale measures from an experimental point of view, e.g. by performing an analysis of their robustness to pool downsampling or of their discriminative power, as well as the exploration of how they behave in statistical significance testing with respect to the traditional evaluation measures which, instead, violate the interval scale assumption.

References

1. Carterette, B.A.: Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. ACM TOIS 30(1), 4:1–4:34 (2012)
2. Ferrante, M., Ferro, N., Pontarollo, S.: A General Theory of IR Evaluation Measures. IEEE TKDE 31(3), 409–422 (March 2019)
3. Kekäläinen, J., Järvelin, K.: Using Graded Relevance Assessments in IR Evaluation. JASIST 53(13), 1120–1129 (November 2002)
4. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of Measurement. Additive and Polynomial Representations, vol. 1. Academic Press, USA (1971)
5. Stevens, S.S.: On the Theory of Scales of Measurement. Science, New Series 103(2684), 677–680 (June 1946)