Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness

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

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

Maria Maistro
Dept. Information Engineering
University of Padua, Italy
maistro@dei.unipd.it

ABSTRACT

In this paper we present a formal framework to define and study the properties of utility-oriented measurements of retrieval effectiveness, like AP, RBP, ERR and many other popular IR evaluation measures. The proposed framework is laid in the wake of the representational theory of measurement, which provides the foundations of the modern theory of measurement in both physical and social sciences, thus contributing to explicitly link IR evaluation to a broader context. The proposed framework is minimal, in the sense that it relies on just one axiom, from which other properties are derived. Finally, it contributes to a better understanding and a clear separation of what issues are due to the inherent problems in comparing systems in terms of retrieval effectiveness and what others are due to the expected numerical properties of a measurement.

Categories and Subject Descriptors

H.3.4 [**Information Search and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance, Theory

Keywords

Representational Theory of Measurement; Omomorphism; Swap; Replacement; Balancing Index

1. INTRODUCTION

Information Retrieval (IR) has been deeply rooted in experimentation since its inception and we often hear quotes like "*To measure is to know*" or "*If you cannot measure, you cannot improve it*", attributed to Sir William Thompson first baron of Kelvin, to remark the importance of experimental evaluation as a means to foster research and innovation in the field.

ICTIR'15, September 27–30, Northampton, MA, USA.
 ISBN 978-1-4503-3833-2/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808194.2809452.

However, even if evaluation has greatly contributed to the advancement of IR, we still lack a deep comprehension about what the evaluation measures we daily employ are and this, somehow, hinders the "*to measure*" part in Lord Kelvin's quotes. This is witnessed by the fact that our understanding of evaluation measures is mostly tied to empirical evidence: for example, we use different kinds of correlation analysis [19, 35] to see how close two evaluation measures are, we adopt different pool downsampling techniques to study the robustness of measures to incomplete information [7, 34], we analyse their sensitivity, stability and discriminative power [6, 28], and so on.

We, as others [3, 8, 16, 26], think that a better comprehension of evaluation measures is needed and that the development of a formal theory to define what an evaluation measure is and to derive and study its properties can be the way to address this need.

In this paper, we start to lay the foundations for a formal framework for utility-oriented measurements of retrieval effectiveness. In particular, we place our work in the broader framework of the *representational theory of measurement* [21], which provides the foundations of the modern theory of measurement in both physical and social sciences.

Our work differs from previous attempts to formalize IR evaluation measures in three main aspects: (i) for the first time, it explicitly puts IR measures in the wake of the measurement theory adopted in other branches of science; (ii) it provides a deeper understanding of what issues are due to the intrinsic difficulties in comparing runs rather than attributing them to the expected numerical properties of a measure; (iii) it is minimal, basically consisting of just one axiom (Definition 5.1), which makes the framework easy and intuitive to grasp and from which the other needed properties are (and will be) derived.

The paper is organized as follows: Section 2 explains the basic concepts of the representational theory of measurement and how our framework will lay on it; Sections 3 to 6 introduce our framework; finally, Section 7 wraps up the discussion and outlooks some future work.

2. MEASUREMENT AND MEASURE

2.1 Representational Theory of Measurement

> **Measurement** is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them accordingly to clearly defined rules [12].

The *representational theory of measurement* [21] aims at providing a formal basis to our intuition about the way the world works. According to the above definition of measurement, the numbers or symbols we collect as measures about the attributes of the entities we examine should be such that their processing and manipulation maintain the relationships among the actual entities under examination in the real world. Therefore, at the basis of measurement, there are the relationships among entities and how we empirically observe them [14].

Consider, for example, the attribute "height" of a tree: in the real world, we are easily able to recognize that some trees are "taller than" others. "Taller than" is an *empirical relation* for height (of a tree) and we can think at it as a *mapping* from the real world to a formal mathematical one, namely from the set of trees to the set of real numbers, provided that, whenever a tree is "taller than" another one, any measure of height assigns a higher number to that tree.

This is the so called *representation condition* which ensures that a measurement must map attributes of entities into numbers (symbols) and empirical relations into numerical (symbolic) ones so that the empirical relations imply and are implied by the numerical (symbolic) ones.

More formally [21, 24], a *relational structure* 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:

- 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;
- 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,

with the condition that $\forall r \in R_X, \forall x_i \in X$, 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.

Note that we talk about a homomorphism rather than an isomorphism because M is generally not one-to-one; in general $M(a) = M(b)$ does not mean that two trees are identical but merely of equal height.

A relational structure $\mathbf{E}$ is called *empirical* if its domain set E spans over the entities under consideration in the real world, e.g. the set of trees; a relational structure $\mathbf{S}$ is called *symbolic* if its domain set S spans over a given set of symbols, e.g. the set of positive real numbers $\mathbb{R}_0^+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

We can now provide a more precise definition of measurement on the basis of the just introduced concepts

> **measurement** is a homomorphism $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ from the real world to a symbolic world. Consequently, a **measure** is the number or symbol assigned to an entity by this mapping in order to characterize an attribute [12].

As an example, consider a set of rods R [21] where an order relation $\preceq$ and a concatenation operation $\circ$ among rods exist. Note that $\preceq$ is a binary relation on the set of rods R while $\circ$ is a ternary one which assigns to each pair of rods a third rod representing their concatenation. Then, the empirical relational structure $\mathbf{E} = \langle A, \preceq, \circ \rangle$ can be mapped into the symbolic relational structure $\mathbf{S} = \langle \mathbb{R}_0^+, \leq, + \rangle$, using as mapping function $M(\cdot)$ the length of a rod so that $a \preceq b \Leftrightarrow M(a) \leq M(b)$ and $M(a \circ b) = M(a) + M(b)$. Note that this example covers also the basics of the classical measure theory [4, 15], where the order relation among sets is given by $A \preceq B \Leftrightarrow A \subseteq B$ and the concatenation operation among two disjoint sets $A \cap B = \varnothing$ is given by $\circ = A \cup B$; a measure is then requested to be *monotonic* $A \subseteq B \Rightarrow M(A) \leq M(B)$ and *additive* $A \cup B \Rightarrow M(A) + M(B)$ when two sets are disjoint $A \cap B = \varnothing$.

2.2 Our Framework

The core of our framework is to start individuating an empirical relational structure $\mathbf{E} = \langle IRS, \preceq \rangle$ which allows us to compare and order different IR systems on the basis of the utility they provide to their users [9, 11, 29]. Clearly, being an empirical relational structure, it is assumed to exist in the real word, i.e. users have their own intuitive notion of when a system is better than another one. In Section 4 we will make this intuitive notion explicit, at least for the cases where it is possible to determine a commonly shared agreement about when a system is better than another one, thus leading to a partial ordering among systems.

We will then individuate a suitable symbolic relational structure $\mathbf{S} = \langle \overline{\mathbb{R}}_0^+, \leq \rangle$ with $\overline{\mathbb{R}}_0^+ = \mathbb{R}_0^+ \cup \{\infty\}$ and, in Section 5, we will provide a definition of IR utility-oriented measurement as a homomorphism between these two relational structures, i.e. we will provide a representation condition. We will also provide an equivalence theorem which allows us to easily verify the representation condition in terms of two simple properties, *swap* and *replacement*, i.e. to check in practice when an evaluation measure like AP or nDCG is actually a measurement in the previous sense. Note that according to the above definition AP or nDCG are called *measurement* while the actual numerical value computed by AP or nDCG for a given run and topic is called *measure*.

Finally, we will also introduce the concept of *balancing* meant to explore the behaviour of a measurement when, in the empirical relational structure, the ordering between two systems is not a priori known. We will show that balancing accounts for the top heaviness of a measurement and we will conduct a preliminary experiment to validate the meaningfulness of its numerical value.

The problem of grounding IR evaluation measures into a broader approach to measuring is a longstanding and crucial one [16]. C. J. van Rijsbergen was early pointing out the issues we encounter with IR evaluation measures [31]:

> In the physical sciences there is usually an empirical ordering of the quantities we wish to measure [...] Such a situation does not hold for information retrieval. There is no empirical ordering for retrieval effectiveness and therefore any measure of retrieval effectiveness will by necessity be artificial

We are not claiming to have fully addressed this hard problem in the present work but rather to have started laying foundations which can contribute to its solution. Moreover, to the best of our knowledge, this is the first attempt

to systematically apply the representational theory of measurement in the context of IR evaluation.

Indeed, [3, 26] stated numerical properties and constraints IR evaluation measures should comply on a case-by-case basis, e.g. when a system retrieves one more relevant document than another one, but they did not build up on an explicit relational structure among systems. [8, 22] built their formal framework for IR evaluation measures on the notion of measurement scale [12, 30], which somehow comes after the definition of measurement; here, we prefer to start from the definition of what IR utility-oriented measurements are and we leave for future work a throughout study of the issues concerning the scales for such measurements. [2] provided a formal framework concerning measures for clustering rather than for IR, even it has been extended in [3] to include also IR measures. Finally, [5] sought for two axioms which allowed him to define when an IR evaluation measure could be expressed as a linear combination of the number of relevant retrieved documents and the number of not-relevant not retrieved documents, which is a different problem from the one of the present paper.

3. PRELIMINARY DEFINITIONS

We stem from [1, 13] for defining the basic concepts of topics, documents, ground-truth, run, and judged run. To the best of our knowledge, these basic concepts have not explicitly defined in previous works [3, 8, 22, 26].

Note that we need to define the same concepts for both set-based retrieval and rank-based retrieval and, to keep the notation compact and stress the similarities between these two cases, we will use the same symbols in both cases – e.g. r_t for run, $D(n)$ for set of retrieved documents by a run, $\mathcal{D}$ for universe set of documents and so on – being clear later on from the context whether we will refer to the set-based or rank-based version.

3.1 Topics, Documents, Ground-truth

Let us consider a set of **documents** D and a set of **topics** T; note that D and T are typically finite sets but we can account also for countable infinite ones.

Let $(REL, \preceq)$ be a totally ordered set of **relevance degrees**, i.e. they are defined on an ordinal scale [30], where we assume the existence of a minimum that we call the **non-relevant** relevance degree $\mathrm{nr} = \min(REL)$. Note that REL is typically a finite set but we can account also for an infinite one. In the former case, we can represent both binary relevance[1] $REL = \{\mathrm{nr}, \mathrm{r}\}$ (non relevant and relevant) and graded relevance [18], e.g. $REL = \{\mathrm{nr}, \mathrm{pr}, \mathrm{hr}\}$ (non-relevant, partially relevant, highly relevant); in the latter case, we can represent both continuous relevance [18] and relevance assigned using unbounded scales, e.g. by using magnitude estimation [23]. Note that the definition of the REL set can accomplish both a notion of "immutable" relevance, as the one somehow adopted in evaluation campaigns, and a notion of relevance dependent on users and their context. In the latter case, we will have different REL sets corresponding to each user/context.

In the following, and without any loss of generality, we consider $REL \subseteq \mathbb{R}_0^+$ with the constraint that $0 \in REL$ and the order relation $\preceq$ becomes the usual ordering $\leq$ on real numbers, which ensures that a higher number corresponds to a higher relevance degree; the non-relevant degree is therefore given by $\min(REL) = 0$. Note that most of the algebraic operations we typically perform on numbers, like addition and multiplication, will be in general senseless on REL, since we take for granted only its order property. As above, this choice allows us to represent the most common cases, i.e. both binary relevance with $REL = \{0, 1\}$ and graded relevance, either discrete with $REL \subseteq \mathbb{N}_0$ or continuous with $REL \subseteq \mathbb{R}_0^+$ in general.

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. Note that, in the case of more complex situations like crowdsourcing for relevance assessment, we can define different GT maps, one for each crowd-worker.

The **recall base** is the map RB from T into $\mathbb{N}$ defined as the total number of relevant documents for a given topic $t \mapsto RB_t = \big|\{d \in D : GT(t, d) > 0\}\big|$. The recall base is a quantity often hard to know in reality and, in some applications, it may be preferable to substitute it with a family of random variables $(t, \omega) \mapsto RB_t(\omega)$ which represents the unknown number of relevant documents present in the collection for every topic, that we will be able at most to estimate. For simplicity, in the sequel we will denote by RB_t the recall base in both the cases, omitting in the latter the dependence on ω.

3.2 Set-based Retrieval

Given a positive natural number n called the *length of the run*, we define the **set of retrieved documents** as $D(n) = \big\{\{d_1, \ldots, d_n\} : d_i \in D\big\}$ and the **universe set of retrieved documents** as $\mathcal{D} := \bigcup_{n=1}^{|D|} D(n) = 2^D$, which is the power set of D, i.e. the set of all the subsets of D.

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

$$t \mapsto r_t = \{d_1, \ldots, d_n\}$$

Note that, since D can be an infinite set, we can have runs retrieving infinite documents.

A multiset (or bag) is a set which may contain the same element several times and its multiplicity of occurrences is relevant [20]. A **set of judged documents** is a (crisp) multiset $(REL,\ m) = \{rel_1, rel_2, rel_1, rel_2, rel_2, rel_4, \ldots\}$, where m is a function from REL into $\overline{\mathbb{N}}_0 = \mathbb{N}_0 \cup \{\infty\}$ representing the multiplicity of every relevance degree rel_j [25]; if the multiplicity is 0, a given relevance degree is simply not present in the multiset, as in the case of rel_3 in the previous example. Note that the multiplicity function m can lead to infinite multisets, when needed. Suppose $\mathcal{M}$ is the infinite set of all the possible multiplicity functions m, then the **universe set of judged documents** is the set $\mathcal{R} := \bigcup_{m \in \mathcal{M}} (REL, m)$ of all the possible sets of judged documents (REL, m).

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

$$(t, r_t) \mapsto \hat{r}_t = \big\{GT(t, d_1), \ldots, GT(t, d_n)\big\} = \big\{\hat{r}_{t,1}, \ldots, \hat{r}_{t,n}\big\}$$

3.3 Rank-based Retrieval

Given a positive natural number n called the *length of the run*, we define the **set of retrieved documents** as

[1]Binary relevance is often thought to be on a categorical scale but, since the scale consists only of two categories one of which indicates the absence of relevance, we can safely consider it as an ordinal scale in fact.

$D(n) = \{(d_1, \ldots, d_n) : d_i \in D, d_i \neq d_j \text{ for any } i \neq j\}$, i.e. the ranked list of retrieved documents without duplicates, and the **universe set of retrieved documents** as $\mathcal{D} := \bigcup_{n=1}^{|D|} D(n)$.

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

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

We denote by $r_t[j]$ the j-th element of the vector r_t, i.e. $r_t[j] = d_j$. Note that, since the cardinality of D may be infinite, we can model also infinite rankings, as those assumed by [27, 33]. We define the **universe set of judged documents** as $\mathcal{R} := \bigcup_{n=1}^{|D|} REL^n$.

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

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

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

4. EMPIRICAL RELATIONAL STRUCTURE

As discussed in Section 2, a key point in defining a measurement is to start from a clear empirical relational structure among the attributes of the entities you would like to measure, in our case the effectiveness of IR systems in terms of the utility they provide to their users [9, 11, 29]. Therefore, $\mathbf{E} = \langle T \times \mathcal{D}, \preceq \rangle$ is our empirical relational structure, i.e. the set of all the runs with an ordering relation where the utility systems provide to their users is roughly expressed in terms of the "amount" of relevance: the more relevance is retrieved by a run, the greater it is.

This is an especially critical point since, as highlighted out by [31], "there is no empirical ordering for retrieval effectiveness". The hardness of this problem clearly emerges also when you consider the actual properties of the set $\mathcal{D}$.

Typically, when you define a measurement, you start from sets having very good properties. For example, in the case of the theory of measure [4, 15], σ-algebras are closed under countable unions, intersections, and complements and the inclusion relation among sets leads to a natural partial ordering. All these nice properties are then reflected in measures and probabilities: since a σ-algebra is closed under countable union, a measure is then requested to be σ-*additive*, i.e. if $\{A_n\}_{n\in\mathbb{N}}$ is a family of disjoints subsets, then $M(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} M(A_n)$ and from this property one obtains that is also *monotone* $A \subseteq B \Rightarrow M(A) \leq M(B)$, since $B = A \cup (A^C \cap B) \Rightarrow M(B) = M(A) + M\big(A \cup (A^C \cap B)\big) \geq M(A)$, which in turn reflects the ordering induced by the inclusion relation on the σ-algebra.

Unfortunately, the set $\mathcal{D}$ lacks many of these desirable properties. For example, inclusion and union on $\mathcal{D}$ would not be as intuitive and agreeable as they are in the case of σ-algebras and this hampers the possibility of requiring additivity as a property of an IR utility-oriented measurement.

Let us consider inclusion: we could say that $r_t \subseteq s_t$ if s_t appends one more document to r_t. Differently from σ-algebras, inclusion would not induce an ordering on $\mathcal{D}$, since you may think that a run retrieving one more relevant document is greater than another one not retrieving it [3, 26] but you may also think that a run retrieving one more not-relevant document is smaller than another one not retrieving it [3], or it should stay equal [26].

The above inclusion can be seen also as a form of union, i.e. as concatenating a run with another one constituted by just a single document, i.e. somehow $s_t = r_t \cup \{d_j\}$. Almost no one would require additivity, i.e. $M(s_t) = M(r_t) + M(d_j)$, and as discussed above there is neither agreement on monotonicity, i.e. when it should be $M(s_t) > M(r_t)$ and when $M(s_t) < M(r_t)$. This is even more evident if you think at data fusion, a kind of much more complicated union: no one would quest for additivity, even in the case of runs without any common document, and consider the performance of the fused run as the sum of the performances of the composing runs, nor they could a priori guarantee monotonicity, ensuring that the performance of the fused run is always greater than or equal to the the performances of the composing runs.

The above mentioned issues with inclusion and union of runs make it difficult also to deal with runs of different length, e.g. constraining the behaviour of a measurement in the symbolic relational structure **S** when runs of different length are somehow contrasted, as it is done in [3, 8, 22, 26], since we basically do not know how to unite and compare them in the empirical relational structure **E**.

Therefore, in this paper, we will focus on a partial ordering among runs of the same length in the empirical relational structure **E**, leading to monotonicity in the symbolic relational structure **S**, and we leave for future work a deeper investigation of inclusion, union, additivity and their implications. In particular, we will restrict ourselves only to those cases where the ordering is intuitive and it is possible to find a commonly shared agreement. Examples of very basic cases are: a run retrieving a relevant document in the first rank position is greater than another one retrieving it in the second position or a run retrieving a more relevant document in a given rank position is greater than another one retrieving a less relevant document in the same position.

The above discussion points out one key contribution of this paper, i.e. highlighting that the core problem in defining an IR measurement is not to constraint its numerical properties (symbolic world) but rather our quite limited understanding of the operations and relationships among runs (empirical world). Indeed, if we better clarify how runs behave in the empirical relational structure, a measurement, intended as a homomorphism between the empirical and symbolic worlds, has to comply with them by construction.

Note that this vision is somehow implicitly present in [8, 22]. Their framework is based on the idea that there must be an agreement between two distinct "relevance measurements", one made by assessors and the other by systems, i.e. how assessors and systems rank documents on the basis of their relevance to a query. Then, they constrain what they call "metric" to the behaviour of the similarity between these two "relevance measurements", but without actually defining what this similarity is . In relation to our work, we could say that the assessor and system "relevance measurements" may somehow resemble the notion of relational structures in the empirical world and the "metric" may in some way approximate the notion of measurement as homomorphism between empirical and symbolic worlds. However, we think that framing the problem in the context of the representational theory of measurement provides more advantages than an ad-hoc approach: it streamlines the core concepts, helps to discuss and address issues at the proper level ei-

ther in the empirical or symbolic worlds, and better links IR evaluation to other sciences. Moreover, we provide an actual partial ordering among runs in the empirical world, from which we derive properties for a measurement, while the concept of similarity is not actually defined by [8, 22].

4.1 Set-based Retrieval

Let us consider two runs r_t and s_t with the same length n. We introduce a **partial ordering among runs** as

$$r_t \preceq s_t \;\Leftrightarrow \big|\{j : \hat{r}_{t,j} \geq rel\}\big| \leq \big|\{j : \hat{s}_{t,j} \geq rel\}\big| \;\; \forall rel \in REL$$

which counts, for each relevance degree, how many items there are above that relevance degree and, if a run has higher counts for each relevance degree, it is considered greater than another one.

For example, if we have four relevance degrees $REL = \{0, 1, 2, 3\}$, the run $\hat{r}_t = \{0, 1, 1, 2, 2\}$ is smaller than the run $\hat{s}_t = \{0, 1, 1, 2, 3\}$ but the run $\hat{r}_t = \{0, 1, 1, 2, 2\}$ is not comparable to the run $\hat{w}_t = \{0, 1, 1, 1, 3\}$ because, relying just on an ordinal scale for the relevance degrees, it is not a priori known whether the decrease from a document with relevance degree 2 to one with relevance degree 1 is compensated or not by the increase from a document with relevance degree 2 to one with relevance degree 3, actually we cannot even say if the two runs are equal.

If we have the relevance grades $REL = \{0, 1, \cdots, q\}$, among all the runs with a fixed number of relevant documents, the run $\{1, \ldots, 1, 0, \ldots, 0\}$ is the smallest, while $\{q, \ldots, q, 0, \ldots, 0\}$ is the greatest one.

In the case of binary relevance, i.e. $REL = \{0, 1\}$, we obtain an intuitive total ordering

$$r_t \preceq s_t \;\Leftrightarrow \big|\{j : \hat{r}_{t,j} \geq 1\}\big| \leq \big|\{j : \hat{s}_{t,j} \geq 1\}\big|$$

where r_t is less than s_t if it retrieved less relevant documents than s_t.

If REL relies on a more powerful scale, e.g. a ratio scale where we can know, for example, that a highly relevant document is twice as relevant as a partially relevant one, the above definition becomes a total ordering also in the case of graded relevance, by basically summing up how many "relevance units" there are in each run.

4.2 Rank-based Retrieval

Let us consider two runs r_t and s_t with the same length n. We introduce a **partial ordering among runs** as

$$r_t \preceq s_t \;\Leftrightarrow \big|\{j \leq k : \hat{r}_t[j] \geq rel\}\big| \leq \big|\{j \leq k : \hat{s}_t[k] \geq rel\}\big| \quad \forall rel \in REL \;\text{ and }\; k \in \{1, \ldots, n\}$$

which counts, for each relevance degree and rank position, how many items there are above that relevance degree and, if a run has higher counts for each relevance degree and rank position, it is considered greater than another one.

For example, if we have four relevance degrees $REL = \{0, 1, 2, 3\}$, the run $\hat{r}_t = (0, 1, 1, 2, 2)$ is smaller than the run $\hat{s}_t = (0, 1, 1, 2, 3)$ but the run $\hat{r}_t = (0, 1, 1, 2, 2)$ is not comparable to the run $\hat{w}_t = (0, 1, 1, 1, 3)$ because, relying just on an ordinal scale for the relevance degrees, it is not a priori known whether the decrease from a document with relevance degree 2 to one with relevance degree 1 at rank 4 is compensated or not by the increase from a document with relevance degree 2 to one with relevance degree 3 at rank 5, as it happens in the set-based retrieval case. On the other hand, the run $\hat{r}_t = (0, 1, 1, 2, 2)$ is not comparable with the run $\hat{v}_t = (2, 0, 1, 2, 1)$ because, even if the document with relevance degree 2 moves forward from rank 5 to rank 1, the backward movement of the document with relevance degree 1 from rank 2 to rank 5 may or may not compensate for it. This latter case points out the effect of ranking with respect to the previous case of set-based retrieval, which would have considered these two runs as equal.

Note that in rank-based retrieval, we cannot achieve a total ordering even in the case of binary relevance. Indeed the run $\hat{r}_t = (0, 1, 0, 1, 0)$ is not comparable to the run $\hat{s}_t = (1, 0, 0, 0, 1)$ because you cannot a priori say whether the forward movement of the relevant document from rank 2 to rank 1 is compensated or not by the backward movement of the relevant document from rank 4 to rank 5.

A possible segmentation of all the runs can be performed in terms of the total number of relevant documents, where a minimum and maximum run can be found. Taking for simplicity $REL = \{0, 1, \ldots, q\}$ and considering a run r_t retrieving just one relevant document, we have that it lays between the minimum and maximum below:

$$(0, \ldots, 0, 1) \preceq \hat{r}_t \preceq (q, 0, \ldots, 0)$$

More in general, for any run r_t retrieving k relevant documents, it holds:

$$(0, \ldots, 0, 1, \ldots, 1) \preceq \hat{r}_t \preceq (q, \ldots, q, 0, \ldots, 0) \qquad (1)$$

Summing up, differently from the case of set-based retrieval, this partial ordering cannot become a total order neither in the case of binary relevance nor in the case of relevance degrees on more powerful scales, e.g. ratio ones. Indeed, the presence of the ranking adds a further dimension which makes impossible to compare every run pair because it is not a priori known how much each rank position influences the ordering.

5. UTILITY-ORIENTED MEASUREMENTS OF RETRIEVAL EFFECTIVENESS

We define an **utility-oriented measurement of retrieval effectiveness** as an homomorphism between the empirical relational structure $\mathbf{E} = \langle T \times \mathcal{D}, \preceq \rangle$, discussed in the previous section, and the symbolic relational structure $\mathbf{S} = \langle \overline{\mathbb{R}}_0^+, \leq \rangle$, that is a mapping which assigns to any sequence of documents $D(n)$ retrieved by a system for a given topic t, a non negative number, i.e. a **utility-oriented measure of retrieval effectiveness**.

More in detail, a utility-oriented measurement of retrieval effectiveness is the composition of a judged run $\hat{r}_t$ with a **scoring function** μ from the universe set of judged documents $\mathcal{R}$ into $\overline{\mathbb{R}}_0^+$ which assigns to any sequence of judged documents a non negative number, ensuring that the ordering $\preceq$ among the runs is properly mapped in the ordering $\leq$ among real numbers.

Definition 5.1. *A function*

$$M : T \times \mathcal{D} \to \overline{\mathbb{R}}_0^+$$

*defined as $M = \mu(\hat{r}_t)$, i.e. the composition of a judged run $\hat{r}_t$ with a scoring function $\mu : \mathcal{R} \to \overline{\mathbb{R}}_0^+$ is a **utility-oriented measurement of retrieval effectiveness** if and only if for any two runs r_t and s_t with the same length n such that $r_t \preceq s_t$, then $\mu(\hat{r}_t) \leq \mu(\hat{s}_t)$.*

Any utility-oriented measurement of retrieval effectiveness is indeed the specification of the scoring function μ and that the property which ensures a proper mapping between the empirical and symbolic relational structures is the *monotonicity* of μ. In this respect, an utility-oriented measurement of retrieval effectiveness is not a "measure" in the classical sense of the measure theory [4, 15], since it lacks the additivity property, but shares with fuzzy measures [32] the fact of relying just on monotonicity.

Note that the monotonicity requested in the definition above differs from the notion of monotonicity in [26], since this latter one applies to runs of different length, which is not our case for the motivations we discussed in the previous section. Similar considerations hold for the notion of document/query monotonicity in [22] which applies to unions of documents/queries.

Even if the previous definition fits our purposes, it could be difficult to check it in practice. Therefore, we introduce two "monotonicity-like" properties, called **replacement** and **swap**, which we will prove to be equivalent to the required monotonicity but are easier to check.

Replacement If we replace a less relevant document with a more relevant one in the same rank position, a utility-oriented measurement of retrieval effectiveness should not decrease. More formally, if

$$r_t = (d_1, \ldots, d_{i-1}, \mathbf{d_i}, d_{i+1}, \ldots, d_n)$$

and

$$s_t = (d_1, \ldots, d_{i-1}, \tilde{\mathbf{d}}_\mathbf{i}, d_{i+1}, \ldots, d_n)$$

with $\mathbf{d_i} \neq \tilde{\mathbf{d}}_\mathbf{i}$ and $\hat{r}_t[i] \leq \hat{s}_t[i]$, then

$$M(r_t) \leq M(s_t)$$

Swap If we swap a less relevant document in a higher rank position with a more relevant one in a lower rank position, a utility-oriented measurement of retrieval effectiveness should not decrease. More formally, if

$$r_t = (d_1, \ldots, d_{i-1}, \mathbf{d_i}, d_{i+1}, \ldots, d_{j-1}, \mathbf{d_j}, d_{j+1}, \ldots, d_n)$$

and

$$s_t = (d_1, \ldots, d_{i-1}, \mathbf{d_j}, d_{i+1}, \ldots, d_{j-1}, \mathbf{d_i}, d_{j+1}, \ldots, d_n)$$

with $\hat{r}[i] \leq \hat{r}[j]$, then

$$M(r_t) \leq M(s_t)$$

The above definitions of replacement and swap are formulated in the case of rank-based retrieval; clearly, in the set-based retrieval case only replacement makes sense while swap does not apply since there is no ranking among documents.

Note that the swap property somehow recalls the idea of priority constraint in [3] and of convergence in [26].

THEOREM 5.2 (EQUIVALENCE). *A scoring function μ defined from $\mathcal{R}$ into $\overline{\mathbb{R}}_0^+$ leads to a utility-oriented measurement of retrieval effectiveness M if and only if it satisfies the Replacement and the Swap properties.*

PROOF. : If μ leads to a utility-oriented measurement of retrieval effectiveness, the Replacement property is clearly a special case of the monotonicity of μ.

Let us now define

$$A(r_t, k, p) = |\{i \leq k : \hat{r}_t[i] \geq p\}|$$

and assume that

$$r_t = (d_1, \ldots, d_{i-1}, \mathbf{d_i}, d_{i+1}, \ldots, d_{j-1}, \mathbf{d_j}, d_{j+1}, \ldots, d_n)$$

and

$$s_t = (d_1, \ldots, d_{i-1}, \mathbf{d_j}, d_{i+1}, \ldots, d_{j-1}, \mathbf{d_i}, d_{j+1}, \ldots, d_n)$$

with $\hat{r}_t[i] \leq \hat{r}_t[j]$.

It is clear that $A(r_t, k, p) = A(s_t, k, p)$ for any $k \leq i-1$ and $p \in \mathbb{R}_0^+$. If $k = i, i+1, \ldots, j-1$, we have $A(r_t, k, p) = A(s_t, k, p)$ for $p < \hat{r}_t[j]$ and $A(r_t, k, p) < A(s_t, k, p)$ for $p \geq \hat{r}_t[j]$, while for $k > j$ again $A(r_t, k, p) = A(s_t, k, p)$ for any $p \in \mathbb{R}_0^+$. This implies that $r_t \preceq s_t$: by the monotonicity we get that $\mu(\hat{r}_t) \leq \mu(\hat{s}_t)$ and the Swap property is proved.

Let us now assume that the Replacement and the Swap properties are satisfied by M. Taken $r_t \preceq s_t$, our aim is to prove that we are able to construct an increasing sequence of runs

$$r_t = r_t^0 \preceq r_t^1 \preceq r_t^2 \preceq \ldots \preceq r_t^h = s_t$$

such that $\mu(\hat{r}_t^j) \leq \mu(\hat{r}_t^{j+1})$ for any $j = 0, \ldots, h-1$, which proves the monotonicity of μ. Let us start from the last term in both the collections of judged runs. If $\hat{r}_t[n] = \hat{s}_t[n]$, we define $r_t^1 = r_t$ and pass to the $n-1$-th element. If $\hat{r}_t[n] < \hat{s}_t[n]$, we replace the last document in r_t with a document of relevance degree $\hat{s}_t[n]$ and define this new run as r_t^1. We have that $r_t^0 = r_t \preceq r_t^1$, by the replacement that $\mu(\hat{r}_t^0) \leq \mu(\hat{r}_t^1)$ and we pass to consider the $n-1$-th element. If $\hat{r}_t[n] > \hat{s}_t[n]$, we swap the last document in r_t with the closest document of minimum relevance grade of the same run. For example, if

$$\hat{r}_t = (1, 0, 1, 0, 1, 1) \quad \text{and} \quad \hat{s}_t = (1, 1, 0, 1, 1, 0)$$

we define $\hat{r}_t^1 = (1, 0, 1, 1, 1, 0)$. It is immediate to see that the new last element of r_t has a relevance degree smaller than or equal to $\hat{s}_t[n]$. Indeed, if on the contrary we assume that $\hat{r}_t[k] > \hat{s}_t[n]$ for any $k < n$ and we define $p = \min\{\hat{r}_t[i], 0 \leq i \leq n\}$, we have that

$$A(r_t, n, p) > A(s_t, n, p)$$

which is in contradiction with the hypothesis that $r_t \preceq s_t$. We have that $r_t^0 = r_t \preceq r_t^1$ and by the swap property that $\mu(\hat{r}_t^0) \leq \mu(\hat{r}_t^1)$. Proceeding now as before in the case that $\hat{r}_t^1[n] = \hat{s}_t[n]$ or $\hat{r}_t^1[n] < \hat{s}_t[n]$, we (possibly) define a new run r_t^2 such that $r_t^1 \preceq r_t^2$ and we pass to consider the $n-1$-th element. Repeating this procedure to the $n-1$-th element, the $n-2$-th element and so on we construct the desired sequence of runs and the monotonicity is proved. □

The same theorem can be proved in the case of set-based retrieval by using just the Replacement property.

As a final remark, note that for any two runs r_t and s_t such that $r_t \preceq s_t$, Definition 5.1 ensures that any two utility-oriented measurements M_1 and M_2 will order r_t below s_t, i.e. $M_1(r_t) \leq M_1(s_t)$ and $M_2(r_t) \leq M_2(s_t)$. On the contrary, when two runs are not comparable, i.e. when they are outside the partial ordering $\preceq$ and we cannot say which one is the greater, we can find two utility-oriented measurements M_1 and M_2 which order them differently.

Consider, for example the following runs

$$r_t = (1, 0, 0, 1, 0) \quad \text{and} \quad s_t = (0, 1, 1, 0, 1),$$

We obtain that

$$Prec(r_t)[5] = \frac{2}{5} < Prec(s_t)[5] = \frac{3}{5}$$

while

$$AP(r_t) = \frac{1}{RB_t}\frac{3}{2} > AP(s_t) = \frac{1}{RB_t}\frac{53}{30}$$

Therefore, Precision judges preferable s_t, while *Average Precision (AP)* r_t.

5.1 Examples of Application of the Equivalence Theorem

In this section, we use the equivalence Theorem 5.2 to show how to demonstrate that an existing IR evaluation measure is an utility-oriented measurements of retrieval effectiveness.

The proof is trivial in the case of *Average Precision (AP)*, *Rank-Biased Precision (RBP)* [27], and *Normalized Discounted Cumulated Gain (nDCG)* [17] and not reported here due to space reasons. Here, we present the case of *Expected Reciprocal Rank (ERR)* [10], which is more interesting.

Given a run r_t of length n, the ERR is defined as

$$ERR(x_1,\dots,x_n) = \sum_{i=1}^{n}\frac{1}{i}\prod_{k=1}^{i-1}(1-x_k)x_i$$

with the convention that $\prod_{i=1}^{0} = 1$ and x_i represents the probability that a user leaves his search after considering the document at position i. An additional assumption is that the map $\hat{r}_t[i] \mapsto x_i(\hat{r}_t[i])$ is increasing and $x_i(0) = 0$.

Let us consider the **Replacement property** and to avoid trivial cases, take $\hat{r}_t[i] < \hat{s}_t[i]$. The property is satisfied if the function $(x_1,\dots,x_n) \mapsto ERR(x_1,\dots,x_n)$ is non-decreasing in any variable. With this aim, we will prove that the partial derivatives $\frac{\partial}{\partial x_k}ERR > 0$ for any $k \le n$ and $(x_1,\dots,x_n) \in [0,1]^n$. It is immediate that $\frac{\partial}{\partial x_n}ERR = \frac{1}{n}\prod_{k=1}^{n-1}(1-x_k) > 0$. Let us now consider $\frac{\partial}{\partial x_{n-1}}ERR$. Denoting $A(x_i,\dots,x_j) = \prod_{k=i}^{j}(1-x_k)$, we get $\frac{\partial}{\partial x_{n-1}}ERR = A(x_1,\dots,x_{n-2})\Big(\frac{1}{n-1} - \frac{x_n}{n}\Big) > 0$ since $\frac{1}{n-1} - \frac{x_n}{n} > \frac{1}{n-1} - \frac{1}{n} > \frac{1}{(n-1)n} > 0$.

The general case follows similarly: take $k < n-1$ and consider $\frac{\partial}{\partial x_k}ERR$. This partial derivative will be positive if and only if

$$S(x_{k+1},\dots,x_n) = \frac{1}{k} - \frac{1}{k+1}x_{k+1}$$

$$-\frac{1}{k+2}A(x_{k+1})x_{k+2} - \dots - \frac{1}{n}A(x_{k+1},\dots,x_{n-1})x_n > 0.$$

Considering the last two terms, we get

$$\frac{1}{n-1}A(x_{k+1},\dots,x_{n-2})x_{n-1} + \frac{1}{n}A(x_{k+1},\dots,x_{n-1})x_n$$

$$\le A(x_{k+1},\dots,x_{n-2})\frac{1}{n-1}\ .$$

This implies that

$$S(x_{k+1},\dots,x_n) > \frac{1}{k} - \dots - \frac{1}{n-2}A(x_{k+1},\dots,x_{n-3})x_{n-2}$$

$$-\frac{1}{n-1}A(x_{k+1},\dots,x_{n-2})$$

Applying the previous computation with the new last two terms and repeating this procedure on and on, at the end we obtain that

$$S(x_{k+1},\dots,x_n) > \frac{1}{k} - \frac{1}{k+1} > 0$$

and the replacement is proved for ERR.

The **Swap property** is a little more challenging. We have

$$ERR = F(x_1,\dots,x_{i-1}) + \frac{1}{i}\prod_{k=1}^{i-1}(1-x_k)\mathbf{x_i}$$

$$+\frac{1}{i+1}\prod_{k=1}^{i-1}(1-x_k)(1-\mathbf{x_i})x_{i+1} + \dots$$

$$\dots + \frac{1}{j-1}\prod_{k=1}^{i-1}(1-x_k)(1-\mathbf{x_i})(1-x_{i+1})\cdots(1-x_{j-2})x_{j-1} +$$

$$+\frac{1}{j}\prod_{k=1}^{i-1}(1-x_k)(1-\mathbf{x_i})\cdots(1-x_{j-1})\mathbf{x_j} + G(x_1,\dots,x_n)\ ,$$

where F and G are suitable functions, while $ERR(s)$ has the same expression with the $\mathbf{x_i}$'s and $\mathbf{x_j}$'s interchanged. It is immediate that $ERR(r_t) \le ERR(s_t)$ if $j = i+1$. Indeed, we have that the previous inequality holds if and only if

$$\frac{1}{i}\mathbf{x_i} + \frac{1}{i+1}(1-\mathbf{x_i})\mathbf{x_{i+1}} \le \frac{1}{i}\mathbf{x_{i+1}} + \frac{1}{i+1}(1-\mathbf{x_{i+1}})\mathbf{x_i}$$

which is equivalent to $\frac{1}{i(i+1)}\mathbf{x_i} \le \frac{1}{i(i+1)}\mathbf{x_{i+1}}$. If $|i-j| > 1$, $ERR(r_t) \le ERR(s_t)$ if and only if

$$\mathbf{x_i}D(x_{i+1},\dots,x_{j-1}) \le \mathbf{x_j}D(x_{i+1},\dots,x_{j-1})$$

where

$$D(x_{i+1},\dots,x_{j-1}) = \frac{1}{i} - \frac{1}{i+1}x_{i+1} - \frac{1}{i+2}(1-x_{i+1})x_{i+2} - \dots$$

$$\dots - \frac{1}{j-1}(1-x_{i+1})\cdots(1-x_{j-2})x_{j-1}$$

$$-\frac{1}{j}(1-x_{i+1})\cdots(1-x_{j-2})(1-x_{j-1})$$

It will be therefore sufficient to prove that $D(x_1,\dots,x_k) > 0$ for any $(x_1,\dots,x_k) \in [0,1]^k$, where $k = j-i-1 > 0$. Let us prove this by induction on k: if $k = 1$ we get

$$D(x_1) = \frac{1}{i} - \frac{x_1}{i+1} - \frac{(1-x_1)}{i+2} \ge \frac{1}{i(i+1)}$$

for any $x_1 \in [0,1]$. Let us now assume that $D(x_1,\dots,x_i) > 0$ for any $i \le k-1$ and $(x_1,\dots,x_i) \in [0,1]^i$. It holds

$$D(x_1,\dots,x_k) = \quad D(x_1,\dots,x_{k-1})$$

$$+\frac{1}{(i+k-1)(i+k)}(1-x_1)\cdots(1-x_{k-1}) > 0$$

for any $(x_1,\dots,x_k) \in [0,1]^k$ and the property is proved.

6. BALANCING

In this section, we explore the behaviour of utility-oriented measurements when two runs r_t and s_t are not comparable according to the the partial ordering $\preceq$.

Let n be the length of a run, let r_t and s_t be two runs, $q_{min} = \min\{rel \in REL : rel > 0\}$ be the minimum relevance degree above not relevant and $q_{max} = \max\{rel \in REL\}$ be the maximum relevance degree, and $M(\cdot)$ a utility-oriented measurement. We assume here that $0 < q_{min} \leq q_{max} < \infty$.

We define the **Balancing Index** as

$$B(n) = \max\Big\{b \in \mathbb{N}\colon M\big(r_t : \hat{r}_t[1] = q_{max}, \hat{r}_t[j] = 0,\ 1 < j \leq n\big) \leq M\big(s_t : \hat{s}_t[i] = 0,\ 1 \leq i < b, \hat{s}_t[j] = q_{min},\ b \leq j \leq n\big)\Big\}$$

As an example, let us consider the case of four relevance degrees $REL = \{0, 1, 2, 3\}$ and runs of length 5. The balancing index seeks the maximum rank position b for which $M\big((3,0,0,0,0)\big)$ is balanced by $M\big((0,0,0,0,1)\big)$ or $M\big((0,0,0,1,1)\big)$ or $M\big((0,0,1,1,1)\big)$ or $M\big((0,1,1,1,1)\big)$, i.e. it determines when the greatest run possible with just one maximally relevant document (3 in this case) is scored "the same" as the smallest run possible with an increasing number of minimally relevant documents (1 in this case).

The balancing index exploits the Replacement and Swap properties in a way, different from the one used in the equivalence theorem, that allows us to move among runs not comparable for the empirical ordering $\preceq$.

In the above example, we have that

$$(3,0,0,0,0) \xrightarrow[\succeq]{\text{Swap}} (0,0,0,0,3) \xrightarrow[\succeq]{\text{Replacement}} (0,0,0,0,1) \xrightarrow[\preceq]{\text{Replacement}} (0,0,0,1,1) \xrightarrow[\preceq]{\text{Replacement}} (0,0,1,1,1) \xrightarrow[\preceq]{\text{Replacement}} (0,1,1,1,1)$$

where every two adjacent run pairs in the chain are comparable according to the empirical ordering $\preceq$ but not the first run with the last ones, e.g. $(3,0,0,0,0)$ is not a priori comparable to $(0,0,0,1,1)$ because neither you know whether the loss of a document with relevance degree 3 is compensated or not by two documents with relevance degree 1 nor you know the effect of ranking.

The balancing index allows us to explore cases that fall outside the empirical ordering $\preceq$ and to characterize the behaviour of the measurements in those circumstances where Definition 5.1 cannot ensure they will a priori act in a homogeneous way.

In particular, a measurement with $B(n) \to n$ behaves like a binary set-based measure, being extremely sensitive to the presence of additional relevant documents in the lowest ranks. On the contrary, a measurement with $B(n) \to 1$ is not sensitive to the presence of additional relevant documents after a relevant one in the top rank.

The balancing index models the concept of *top heaviness*, an important and somehow desired characteristic of a measurement, as highlighted also in previous works. The closeness threshold constraint [3] resembles it, even if it is formulated as a constraint stating that relevant documents in top ranks should count more rather than as an index you can actually compute to characterize a measurement; similar considerations hold for the notion of top-weightedness [26]. However, it should be noted that, instead of requesting top heaviness to be an a-priori propriety as in [3, 26], the balancing index explicitly points out that top heaviness is a property of the measurements that concerns the area where runs are not a priori comparable, i.e. outside the empirical ordering $\preceq$, and this, in turn, causes measurements to possibly behave differently one from another, being more or less top heavy.

With respect to other empirical indexes for quantifying top heaviness, the balancing index has the advantage that it can be derived analytically. Below, some example of balancing indexes for some popular measurements are reported:

AP

$$B(n) = \max\left\{b \in \mathbb{N}\colon \sum_{k=1}^{n-b+1} \frac{k}{k+b-1} \geq 1\right\}$$

RBP

$$B(n) = \max\left\{b \in \mathbb{N}\colon b \geq \log_p(1-p+p^n)+1\right\}$$

where p is the persistence parameter of RBP.

ERR

$$B(n) = \max\left\{b \in \mathbb{N}\colon x_{min} \sum_{k=b}^{n} \frac{(1-x_{min})^{k-1}}{k} \geq x_{max}\right\}$$

where x_{min} represents the probability that a user leaves his search after considering a document of relevance q_{min} and x_{max} represents the probability that a user leaves his search after considering a document of relevance q_{max}.

nDCG

$$B(n) = \max\{b_1, b_2\}$$

where

$$b_1 = \max\left\{b > a \in \mathbb{N}\colon \sum_{k=0}^{n-b} \frac{q_{min}}{log_a(k+b)} \geq q_{max}\right\},$$

$$b_2 = \max\left\{b \leq a \in \mathbb{N}\colon (a-b+1)q_{min} + c \geq q_{max}\right\},$$

$$c = \sum_{k=0}^{n-a-1} \frac{q_{min}}{log_a(k+a+1)}$$

and a is the base of the logarithm in nDCG. Recall that $\max\{\varnothing\} = -\infty$.

It can be noted that some of the above formulas depend explicitly on the length of the run under consideration, as in the case of RBP, while others have an implicit dependence on it and might be more complex to be computed.

Therefore, we defined an algorithm which allows us to compute the balancing index numerically. The complexity of the algorithm is $O(n)$, since, assuming that the computation of the measurement M requires a constant number of operations, the while loop carries out at most $n-1$ iterations and at any iterations it performs a constant number of operations.

Note that, even if we compute the balancing index in a numerical way, it is not an empirical indicator, as for example the discriminative power [28] is, whose computation depends on a given experimental collection and a set of runs and whose value may change from dataset to dataset.

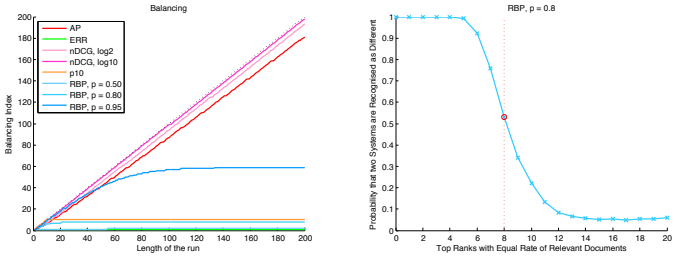
Figure 1.(a) reports the balancing index for several evaluation measurements at different run lengths. It can be noted that for AP and nDCG we have $B(n) \to n$ since it is close

Algorithm 1 Algorithm for computing the balancing index.

Require: n, the length of the run; q_{min} and q_{max} the minimal and maximal relevance degrees

Ensure: b, the balancing index for a run of length n

procedure BALANCING(n, q_{min}, q_{max})
  $refValue \leftarrow M(r : \hat{r}_t[1] = q_{max}, \ldots$
    $\hat{r}_t[j] = 0 \;\; 1 < j \leq n)$
  $cmpValue \leftarrow M(r : \hat{r}_t[j] = 0 \; 1 \leq j < n, \ldots$
    $\hat{r}_t[n] = q_{min})$
  $b \leftarrow n$
  while $refValue > cmpValue$ **do**
    $b{-}{-}$
    $cmpValue \leftarrow M(r : \hat{r}_t[j] = 0 \; 1 \leq j < b, \ldots$
      $\hat{r}_t[j] = q_{min} \; b \leq j \leq n)$
  end while
  return b
end procedure



(a) Balancing index for AP, RBP, ERR, P10, nDCG for different run lengths.

(b) Test of the meaningfulness of the balancing index for RBP with p = 0.8.

Figure 1: (a) Balancing index for various measures and (b) its evaluation for RBP.

to the bisector, indicating that they are not strongly top-heavy measurements and that they are sensitive to relevant documents in the lower ranks. On the other hand, ERR is the most top-heavy measurement since its balancing index is $b = 1$ for any run length, meaning that missing a relevant document in the first rank position can not be compensated even by a run filled in with relevant documents from the second rank position to the end. RBP falls somehow in-between, still being a quite top-heavy measurement; it can be noted as for $p = 0.8$ the balancing index saturates to $b = 8$ for run lengths greater than 20 while, as p increases, it tends to be less top-heavy with almost $b = 60$ for $p = 0.95$.

In order to assess the meaningfulness of the balancing index, we conducted a preliminary experiment with RBP and $p = 0.8$. We simulated two runs of length $n = 1000$ consisting of 50 topics each, generated as shown in Figure 2.

In the top ranks up to rnk they have the same proportion (20%) of relevant documents; in the ranks from rnk to 20 they have different proportions of relevant documents 70% for r_t and 30% for s_t; in the ranks from 21 to $n = 1000$ they have still different proportions of relevant documents 10% for r_t and 70% for s_t. Then, we increased rnk from 0 to 20: when $rnk = 0$, r_t contains more than twice relevant documents in the top ranks than s_t and much less relevant documents in the very long tail; when $rnk = 20$, r_t and s_t have the same proportion of relevant documents in the top ranks but r_t has much less relevant documents than s_t in all the other rank positions. For each increasing value of rnk, we performed a Student's t test with $\alpha = 0.05$ to assess whether r_t and s_t were significantly different. We repeated this experiment 10,000 times and, for each value of rnk, we computed the probability that the two runs are considered significantly different as the ratio among the number of times the Student's t test rejects the null hypothesis and 10,000, the total number of trials.

Figure 1.(b) shows the results of this experiment. It can be noted that, as far as rnk grows up the balancing index $b = 8$, the fact that r_t contains a bigger proportion of relevant documents than s_t in the top ranks almost always leads to consider the two runs as significantly different. On the other hand, as soon as rnk passes the balancing index $b = 8$ and the proportion of relevant documents in the top ranks of r_t and s_t starts to get more and more similar, the probability of considering the two runs significantly different gets lower and lower, completely ignoring the long tail where they are actually quite different. This is a clear indicator of top-heaviness, well reflected by the balancing index.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have laid the foundations of a formal framework for defining what a utility-oriented measurement of retrieval effectiveness is, on the basis of the representational theory of measurement, putting IR evaluation in the wake of other physical and social sciences as far as measuring is concerned. A core contribution of the paper is to address the problem by clearly separating what are the issues in dealing with comparable/not comparable runs in the empirical world from what are the expected properties of a measurement in the symbolic world.

We proposed a minimal definition of measurement, based on just one axiom (Definition 5.1), and provided an equivalence theorem (Theorem 5.2) to check it in practice, as well as examples of its application.

Finally, we proposed the balancing index as an indicator of the top-heaviness of a measurement, providing both formulas and an algorithm to compute it. We have also conducted a preliminary experiment to show that its numerical value is a meaningful indicator of top-heaviness.

Future work will concern a deeper exploration of the core problems such measurements have, as for example additivity. We will also exploit the theory of scales of measurement in order to study the scales actually adopted by common measurements like AP, RBP, ERR, nDCG and others.

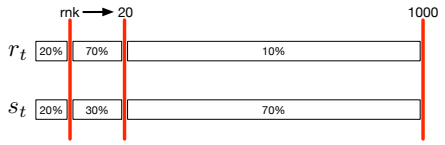


Figure 2: Creation of the simulated runs for assessing the meaningfulness of the balancing index in the case of RBP. Note that the percentages are not referred to the whole run but to each segment separately. Therefore, they do not need to sum up to 100% but to be between 0% and 100% within each segment.

Furthermore, we will consider the application of the proposed framework to other cases, such as measures based on diversity. This will lead to a different definition of the partial ordering $\preceq$ in the empirical relational structure $\mathbf{E}$ to capture the notion of diversity but Definition 5.1 of IR measurement of retrieval effectiveness will remain the same. Moreover, this may also require to individuate properties different from Swap and Replacement to provide an equivalence theorem in the vein of Theorem 5.2 suitable for this case.

8. REFERENCES

[1] M. Angelini, N. Ferro, G. Santucci, and G. Silvello. VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis. *JVLC*, 25(4):394–413, 2014.

[2] E. Amigó, J. Gonzalo, J. Artiles, and M. F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *IR*, 12(4):461–486, 2009.

[3] E. Amigó, J. Gonzalo, and M. F. Verdejo. A General Evaluation Measure for Document Organization Tasks. In SIGIR 2013, pp. 643–652.

[4] P. Billingsley. *Probability and Measure*. John Wiley & Sons, New York, USA, 3rd edition, 1995.

[5] P. Bollman. Two Axioms for Evaluation Measures in Information Retrieval. In SIGIR 1984, pp. 233–245.

[6] C. Buckley and E. M. Voorhees. Evaluating Evaluation Measure Stability. In SIGIR 2000, pp. 33–40.

[7] C. Buckley and E. M. Voorhees. Retrieval Evaluation with Incomplete Information. In SIGIR 2004, pp. 25–32.

[8] L. Busin and S. Mizzaro. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In ICTIR 2013, pp. 22–29.

[9] B. A. Carterette. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In SIGIR 2011, pp. 903–912.

[10] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In CIKM 2009, pp. 621–630.

[11] W. S. Cooper. On Selecting a Measure of Retrieval Effectiveness. *JASIS*, 24(2):87–100, 1973.

[12] N. E. Fenton and J. Bieman. *Software Metrics: A Rigorous & Practical Approach*. Chapman and Hall/CRC, USA, 3rd edition, 2014.

[13] N. Ferro, G. Silvello, H. Keskustalo, A. Pirkola, and K. Järvelin. The Twist Measure for IR Evaluation: Taking User's Effort Into Account. *JASIST*, 2015.

[14] L. Finkelstein. Widely, Strongly and Weakly Defined Measurement. *Measurement*, 34(1):39–48, 2003.

[15] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, USA, 2nd edition, 1999.

[16] N. Fuhr. IR between Science and Engineering, and the Role of Experimentation. In CLEF 2010, p. 1. LNCS 6360.

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

[18] J. Kekäläinen and K. Järvelin. Using Graded Relevance Assessments in IR Evaluation. *JASIST*, 53(13):1120—1129, 2002.

[19] M. G. Kendall. *Rank correlation methods*. Griffin, Oxford, England, 1948.

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

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

[22] E. Maddalena and S. Mizzaro. Axiometrics: Axioms of Information Retrieval Effectiveness Metrics. In EVIA 2014, pp. 17–24.

[23] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. Judging Relevance Using Magnitude Estimation. In ECIR 2015, pp. 215–220. LNCS 9022.

[24] L. Mari. Beyond the Representational Viewpoint: a New Formalization of Measurement. *Measurement*, 27(2):71–84, 2000.

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

[26] A. Moffat. Seven Numeric Properties of Effectiveness Metrics. In AIRS 2013, pp. 1–12. LNCS 8281.

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

[28] T. Sakai. Evaluating Evaluation Metrics based on the Bootstrap. In SIGIR 2006, pp. 525–532.

[29] T. Sakai. Metrics, Statistics, Tests. In *Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Revised Tutorial Lectures*, pp. 116–163. LNCS 8173, 2014.

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

[31] C. J. van Rijsbergen. Retrieval effectiveness. In K. Spärck Jones, editor, *Information Retrieval Experiment*, pp. 32–43. Butterworths, London, United Kingdom, 1981.

[32] Z. Y. Wang and G. J. Klir. *Fuzzy Measure Theory*. Springer-Verlag, New York, USA, 1992.

[33] W. Webber, A. Moffat, and J. Zobel. A Similarity Measure for Indefinite Rankings. *TOIS*, 4(28):20:1–20:38, 2010.

[34] E. Yilmaz and J. A. Aslam. Estimating average precision when judgments are incomplete. *Knowledge and Information Systems*, 16(2):173–211, 2008.

[35] E. Yilmaz, J. A. Aslam, and S. E. Robertson. A New Rank Correlation Coefficient for Information Retrieval. In SIGIR 2008, pp. 587–594.