Basis of a Formal Framework for Information Retrieval Evaluation Measurements*

Marco Ferrante[1], Nicola Ferro[2], and Maria Maistro[2]

[1] Dept. of Mathematics, University of Padua, Italy
ferrante@math.unipd.it
[2] Dept. of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it, maistro@dei.unipd.it

Abstract. In this paper we present a formal framework, based on the representational theory of measurement and we define and study the properties of utility-oriented measurements of retrieval effectiveness like AP, RBP, ERR and many other popular IR evaluation measures.

1 Introduction

Even if *Information Retrieval (IR)* has been deeply rooted in experimentation since its inception, especially with respect to the evaluation area, we still lack a deep comprehension about what the evaluation measures are and we mostly rely just to empirical evidence.

We, as others [1,2,6], think that, in order to achieve a better understanding of evaluation measures, the development of a rigorous theory is needed. Therefore, we start to lay the foundations for a formal framework which differs from previous attempts to formalize IR evaluation measures since it explicitly puts these metrics in the wake of the measurement theory, it distinguishes between issues due to the intrinsic difficulties in comparing runs and those due to the numerical properties of measures, and it is minimal, consisting of just one axiom (Definition 1).

2 Preliminary Definitions

Let us consider a set of **documents** D and a set of **topics** T and let $(REL, \preceq)$ be a totally ordered set of **relevance degrees**, where we assume the existence of a minimum that we call the **non-relevant** relevance degree $\mathrm{nr} = \min(REL)$.

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

* Extended abstract of [4].

Defined 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, and we denote by $\hat{r}_t[j]$ the j-th element of the vector $\hat{r}_t$, i.e. $\hat{r}_t[j] = GT(t, d_j)$, where GT is the **ground-truth**, i.e. a map which assigns the relevance degree.

3 Measurement and Measure

The *representational theory of measurement* [5] aims at providing a formal basis to our intuition concerning the action of measuring. More formally, 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. Given two relational structures $\mathbf{X}$ and $\mathbf{Y}$, a *homomorphism* $\mathbf{M} : \mathbf{X} \to \mathbf{Y}$ is a mapping $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ where $\mathrm{M} : X \to Y$ and $\mathrm{M}_R : R_X \to R_Y$ is a function which maps each relation on the domain set into one 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 $\mathrm{M}_R(r)\Big(\mathrm{M}(x_1), \ldots, \mathrm{M}(x_n)\Big)$. A relational structure $\mathbf{E}$ is called *empirical* if its domain set E spans over the entities in the real world, while it is called *symbolic*, $\mathbf{S}$, if its domain set S spans over a given set of symbols.

Hence, more precisely a **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 [3].

A key point in defining a measurement is to start from a clear empirical relational structure, in our case it is represented by the set of all the runs with an ordering relation based on the utility expressed as "amount" of relevance, $\mathbf{E} = \langle T \times \mathcal{D}, \preceq \rangle$. Since the set $\mathcal{D}$ lacks of many desirable properties, for example, inclusion and union, we will focus on a partial ordering among runs of the same length. 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.

Therefore, let r_t and s_t be two runs with length n, we introduce the above mentioned **partial ordering among runs** as

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

which counts, for each relevance degree and rank position, how many items 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 $\hat{w}_t = (0, 1, 1, 1, 3)$.

4 Utility-oriented Measurements of Retrieval Effectiveness

We define an **utility-oriented measurement of retrieval effectiveness** as an homomorphism between the empirical relational structure $\mathbf{E} = \langle T \times \mathcal{D}, \preceq \rangle$,

and the symbolic relational structure $\mathbf{S} = \langle \overline{\mathbb{R}}_0^+, \leq \rangle$, that is a mapping which assign to any run a non negative number, i.e. a **utility-oriented measure of retrieval effectiveness**.

Definition 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 $\mu(\hat{r}_t) \leq \mu(\hat{s}_t)$.*

Even if the previous definition fits our purposes, it could be difficult to check it in practice. Therefore, we introduce two "monotonicity-like" properties, **replacement** and **swap**, which are equivalent to the required monotonicity (Theorem 1).

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.

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.

Theorem 1 (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.*

As a final remark, note that for any two runs r_t and s_t such that $r_t \preceq s_t$, Definition 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 $\preceq$, we can find two utility-oriented measurements M_1 and M_2 which order them differently.

5 Balancing

In this section, we explore the behaviour of utility-oriented measurements when two runs are not comparable, according to the the partial ordering $\preceq$. Let r_t and s_t be two runs with length n, q_{min} is the minimum relevance degree above not relevant and q_{max} is the maximum relevance degree, hence $0 < q_{min} \leq q_{max} < \infty$. We define the **Balancing Index** as a function of the run length:

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

As an example, consider the binary case $REL = \{0, 1\}$ and runs of length 5. The balancing index seeks the maximum rank position of the first relevant document, for which $M\big((1,0,0,0,0)\big)$ has score greater or equal than $M\big((0,0,0,0,1)\big)$ or $M\big((0,0,0,1,1)\big)$ or $M\big((0,0,1,1,1)\big)$ or $M\big((0,1,1,1,1)\big)$.

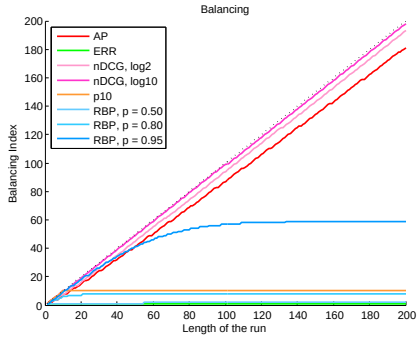

Fig. 1. Balancing index for AP, RBP, ERR, P10, nDCG for different run lengths.

Figure 1 reports the balancing index for several evaluation measurements at different run lengths. It can be noted that *Expected Reciprocal Rank (ERR)* is the most top heavy measurement since $B(n) = 1$, *Average Precision (AP)* and *Normalized Discounted Cumulated Gain (nDCG)* are not strongly top heavy, $B(n) \to n$, while *Rank-Biased Precision (RBP)* falls somehow in-between.

This index explicitly points out that the top heaviness is a property of the measurements that concerns the area where runs are not a priori comparable, and this causes measurements to possibly behave differently one from another, being more or less top heavy.

6 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.

Future work will concern the exploration of measurements core problems, the exploitation of the theory of measurements scales, and the application of the proposed framework to other cases, i.e. measures based on diversity.

References

1. E. Amigó, J. Gonzalo, and M. F. Verdejo. A General Evaluation Measure for Document Organization Tasks. In SIGIR 2013, pp. 643–652.
2. L. Busin and S. Mizzaro. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In ICTIR 2013, pp. 22–29.
3. N. E. Fenton and J. Bieman. *Software Metrics: A Rigorous & Practical Approach*. Chapman and Hall/CRC, USA, 3rd edition, 2014.
4. M. Ferrante, N. Ferro, and M. Maistro. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In ICTIR 2015, pp. 21–30.
5. D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement. Additive and Polynomial Representations*, volume 1. Academic Press, USA, 1971.
6. A. Moffat. Seven Numeric Properties of Effectiveness Metrics. In AIRS 2013, pp. 1–12. LNCS 8281.