Axiometrics: Axioms of Information Retrieval Effectiveness Metrics

Eddy Maddalena
Department of Maths and Computer Science
University of Udine
Udine, Italy
eddy.maddalena@uniud.it

Stefano Mizzaro
Department of Maths and Computer Science
University of Udine
Udine, Italy
mizzaro@uniud.it

ABSTRACT

There are literally dozens (most likely more than one hundred) information retrieval effectiveness metrics, and counting, but a common, general, and formal understanding of their properties is still missing. In this paper we aim at improving and extending the recently published work by Busin and Mizzaro [6]. That paper proposes an axiomatic approach to Information Retrieval (IR) effectiveness metrics, and more in detail: (i) it defines a framework based on the notions of measure, measurement, and similarity; (ii) it provides a general definition of IR effectiveness metric; and (iii) it proposes a set of axioms that every effectiveness metric should satisfy. Here we build on their work and more specifically: we design a different and improved set of axioms, we provide a definition of some common metrics, and we derive some theorems from the axioms.

1. INTRODUCTION

In the Information Retrieval (IR) field, according to survey in 2006 [8], more than 50 effectiveness metrics are identified, taking into account only the system oriented ones. This is a rough underestimate; as discussed for example in [2], more than one hundred IR metrics exist, let alone user-oriented ones or metrics for tasks somehow related to IR, like filtering, clustering, recommendation, summarization, etc. Figure 1 is a graphical representation of this.

This large number is not balanced by a complete understanding of the conceptual and formal properties that any IR effectiveness metrics should satisfy, as discussed also at the recent SWIRL meeting (`http://www.cs.rmit.edu.au/swirl12/`). It is clear that a better understanding of the formal properties of effectiveness metrics would have several advantages, for example it would help to avoid wasting time in tuning retrieval systems according to the wrong metric. However, this is still lacking.

Among some recent attempts to study the formal properties of IR metrics (see next section), in this paper we focus specifically on the work by Busin and Mizzaro [6]. The contribution of that paper is threefold: (i) it defines a framework, grounded on measurement theory, based on the notions of measure, measurement, and similarity; (ii) it provides a general definition of IR effectiveness metric; and (iii) it proposes a set of axioms that every effectiveness metric should satisfy. Although that paper has the merit of proposing to ground on measurement theory to study IR metrics, it also has some limits. First, the analysis of the existing metrics in terms of the framework is very brief, and concerns only four metrics. Therefore, although the framework has been proven adequate to express the axioms, its ability to take into account different metrics has not really been tested. Second, the axioms are a bit disorganized: they are not clearly categorized, and some of them would probably more justifiable as theorems than basic axioms. Third, the usefulnes of the axioms, in terms of deriving theorems from them, is almost not tested, as only one theorem is quickly formalized. We try to overcome those limits in the present work

This paper is structured as follows. In Section 2 we briefly recall the previous work on formal accounts of IR effectiveness metrics. In Section 3 we briefly summarize the framework and notation proposed in [6]. In Section 4 some metrics are defined within the framework, to demonstrate its expressiveness power. In Section 5, a set of axioms, different from that proposed in [6], is stated and the axioms are exploited to derive some theorems in Section 6. Conclusions and future work are presented in Section 7.

2. RELATED WORK

Although formal approaches in the IR field have mainly focussed on the retrieval process rather than on effectiveness metrics themselves (see, e.g., [9, 10]), some research specific to effectiveness metrics does exist, and it is briefly discussed here.

The first early attempts were made by Swets [13], who listed some properties of IR effectiveness metrics, and van Rijsbergen [15, Ch. 7], who followed an axiomatic approach. In [5], Bollmann discusses the risk of obtaining inconsistent evaluations on a document collection and on its subcollections. Two axioms on effectiveness metrics, named the Axiom of monotonicity and the Archimedean axiom, are proposed, and their implications are presented as a theorem. These approaches are developed on the basis of binary relevance (either a document is relevant or it is not) and binary retrieval (either a document is retrieved or it is not). Here we do not make any assumption on the notions of relevance and retrieval (binary, ranked, continuous, etc.). Our approach is meant to be more general.

Yao [17] focusses on the notion of user preferences to measure the relevance (or usefulness) of documents. He adopts a framework where user judgements are described as a weak order. On this basis he then proposes a new effectiveness metric that compares the relative order of documents. The proposed metric is proved to be appropriate through an axiomatic approach.

More recently, Amigó et al. in [1] focus their formal analysis on evaluation metrics for text clustering algorithms find-

Figure 1: IR effectiveness metrics (from [2])

ing four basic formal constraints. These constraints should be intuitive and could point out the limitations of each metric. Moreover, it should be possible to prove formally which constraint is satisfied by a metric or a metric family, not just empirically. They found BCubed metrics (BCubed precision and Bcubed recall) to be the only ones satisfying the four proposed constraints. Amigó et al. in [3] discuss, with a similar approach, a unified comparative view of metrics for document filtering. They discovered that no metric for document filtering can satisfy all desirable properties unless a smoothing process is performed. Finally, in an even more recent work [4], they start from a set of formal constraints to define a general metric for document organization tasks, that include retrieval, clustering, and filtering.

Another attempt is by Moffat, that in [11] lists seven properties of IR metrics: boundedness, monotonicity, convergence, top-weightedness, localization, completeness, and realizability. Several metrics are then analyzed to understand whether they satisfy these properties.

3. THE AXIOMETRICS FRAMEWORK

In the remainder of this paper, we build on the framework recently proposed in [6], based on the notions of measurement, measurement scale, and similarity [16, 12, 14]. Due to space limitations, we can provide only a very short summary; the reader is referred to [6] for further details and motivations.

Given a query q, system and user/assessor relevance measurements on a single document d are denoted by $\sigma(q,d)$ and $\alpha(q,d)$, respectively; the notation is also extended in a straghtforward way to set of documents (D) and queries (Q); and the scale of a measurement (e.g., a binary relevance assessment α) is denoted by $\mathsf{scale}(\alpha) = [\![R,N]\!]$ (where R means relevant and N nonrelevant). Then a notion of similarity between measurements is defined, denoted as

$$\underset{Q,D}{\mathsf{sim}}\left(\alpha(Q,D),\sigma(Q,D)\right)$$

or more briefly as $\mathsf{sim}_{Q,D}(\alpha,\sigma)$ when Q and D are common to both measurements (as it is often the case). In this paper we adopt a different approach from [6], and we will not use the similarity of sets of queries (Q) and/or documents (D): we will need $\mathsf{sim}_{q,d}(\alpha,\sigma)$ only.

In [6] a metric is then defined as a function that takes as arguments two measurements α and σ, a set of documents D, and a set of queries Q, and provides as output a numeric value (usually in $\mathbb{R}$). In mathematical notation, $\mathsf{metric} : \alpha \times \sigma \times D \times Q \mapsto \mathbb{R}$.

Componentwise, a metric can be identified on the basis of five components: $\mathsf{scale}(\alpha)$, $\mathsf{scale}(\sigma)$, a notion of similarity sim, how the values on single documents are averaged over the set D (we denote the corresponding averaging function with avgD), and how these averages are averaged over the set Q (avgQ). We can write:

$$\mathsf{metric}\left(\mathsf{scale}(\alpha),\mathsf{scale}(\sigma),\mathsf{sim},\mathsf{avgD},\mathsf{avgQ}\right). \qquad (1)$$

This framework and notation should be general enough to model most (if not all) effectiveness metrics, as we show next.

4. SOME METRICS

In this section we define the sim, avgD and avgQ functions for some common metrics to demonstrate that the framework and notation should be general enough to model most (if not all) effectiveness metrics.

Table 1 presents tentative definitions of some common metrics. The table should be understandable, but we briefly discuss some metrics. For instance, for both Precision and Recall, $\mathsf{scale}(\alpha) = \mathsf{scale}(\sigma) = [\![R,N]\!]$. Let $Rel = \{d \in D | \alpha(d) = R\}$ and $Ret = \{d \in D | \sigma(d) = R\}$ be the sets of relevant and retrieved documents, respectively. Similarity is (see the table) 1 if a document d is both retrieved and relevant and 0 otherwise. Then, the avgD functions for Precision and Recall are the arithmetic means over the sets Ret and Rel, respectively, and both the avgQ functions are the arithmetic mean over the set Q.

For MAP and MAP-like metrics (GMAP, logitAP, yaAP, etc.) we have two different scales: $\mathsf{scale}(\alpha) = [\![R,N]\!]$ and $\mathsf{scale}(\sigma) = [\![Rank]\!]$. Similarity is more complex on a rank (see the formula in the table), since to understand the similarity of a document d we need to analize also other documents in the rank. Similarity can then be used to define AP values and then MAP is obtained using as avgQ the arithmetic mean of the AP values. GMAP is similar to MAP, the

Metric	scale(α)	scale(σ)	$\mathsf{sim}_{q,d}(\alpha,\sigma)$	avgD	avgQ
Precision				$P_q = \frac{1}{\lvert Ret\rvert}\sum_{d\in Ret}\mathsf{sim}_{q,d}(\alpha,\sigma)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}P_q$
Recall		$[\![R,N]\!]$	$\begin{cases}1 & \text{if } \alpha(d)=\sigma(d)\\ 0 & \text{otherwise}\end{cases}$	$R_q = \frac{1}{\lvert Rel\rvert}\sum_{d\in Rel}\mathsf{sim}_{q,d}(\alpha,\sigma)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}R_q$
P@n				$P@n_q = \frac{1}{n}\sum_{\substack{d\in Rel,\\ \sigma(d)\le n}}\mathsf{sim}_{q,d}(\alpha,\sigma)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}P@n_q$
R-Prec				$R\text{-}Prec_q = \frac{1}{\lvert Rel\rvert}\sum_{\substack{d\in Rel,\\ \sigma(d)\le\lvert Rel\rvert}}\mathsf{sim}_{q,d}(\alpha,\sigma)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}R\text{-}Prec_q$
MAP	$[\![R,N]\!]$		$\begin{cases}1 & \text{if } \alpha(d)=R\\ 1 & \text{if } \alpha(d)=N\wedge\\ & \nexists d' \mid (\alpha(d')=R\wedge\\ & \quad\sigma(d')>\sigma(d))\\ 0 & \text{otherwise}\end{cases}$	$AP_q = \frac{1}{\lvert Rel\rvert}\sum_{d\in Rel}\mathsf{sim}_{q,d}(\alpha,\sigma)\cdot P@\sigma(d)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}AP_q$
MAP@n		$[\![Rank]\!]$		$AP_q = \frac{1}{\lvert Rel\rvert}\sum_{\substack{d\in Rel,\\ \sigma(d)\le n}}\mathsf{sim}_{q,d}(\alpha,\sigma)\cdot P@\sigma(d)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}AP_q$
GMAP				$AP_q = \frac{1}{\lvert Rel\rvert}\sum_{d\in Rel}\mathsf{sim}_{q,d}(\alpha,\sigma)\cdot P@\sigma(d)$	$\sqrt[\lvert Q\rvert]{\prod_{q\in Q}AP_q}$
(log AP)					$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}\log AP_q$
logitAP				$AP_q = \frac{1}{\lvert Rel\rvert}\sum_{d\in Rel}\mathsf{sim}_{q,d}(\alpha,\sigma)\cdot P@\sigma(d)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}\log\frac{AP+\epsilon}{1-AP+\epsilon}$
ADM	$[0,1]$	$[0,1]$	$\lvert\sigma(q,d)-\alpha(q,d)\rvert$	$ADM_q = 1-\frac{1}{\lvert D\rvert}\sum_{d_i\in D}\mathsf{sim}_{q,d_i}(\alpha,\sigma)$	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q}ADM_q$

Table 1: Metrics on the basis of their components as per formula (1).

only difference being on avgQ since in GMAP the geometric mean is used. GMAP can also be defined in an equivalent way as the average of logarithms, and logitAP definition is similar. (and yaAP should be similar as well). The table also includes MAP@n, i.e., MAP computed averaging only the AP values of the relevant documents retrieved in the first n rank positions, and considering 0 as the AP of documents retrieved after rank n (this is the metric used in TREC-like settings). The last row defines ADM [7].

As it is discussed at length in [6], besides allowing us to define the metrics, the framework allows us to state the axioms in a way that is independent from the scales of the relevance measurements.

5. AXIOMS

We now can list some axioms: they define properties that, *ceteris paribus*, any effectiveness metric should satisfy. Axioms can also be interpreted as a set of constraints on a search space. We formalize as axioms the properties of similarity between relevance measurements (Subsection 5.1), we then present some axioms that define the relationships between similarity and metrics (Subsection 5.2), and we then present metric-specific axioms (Subsection 5.3).

5.1 Similarity

The first axioms represent basic constraints on similarity, and metrics are not involved yet.

AXIOM 1 (SIMILARITY OF DOCUMENTS). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurement such that $\alpha(q,d) = \alpha(q,d')$ and $\sigma(q,d) = \sigma(q,d')$. Then*

$$\underset{q,d}{\mathsf{sim}}(\alpha,\sigma) = \underset{q,d'}{\mathsf{sim}}(\alpha,\sigma).$$

AXIOM 2 (SIMILARITY OF QUERIES). *Let q and q' be two queries, d a document, α a human relevance measurement and σ a system relevance measurement such that $\alpha(q,d) = \alpha(q',d)$ and $\sigma(q,d) = \sigma(q',d)$. Then*

$$\underset{q,d}{\mathsf{sim}}(\alpha,\sigma) = \underset{q',d}{\mathsf{sim}}(\alpha,\sigma).$$

AXIOM 3 (SIMILARITY OF TWO SYSTEMS). *Let q be a query, d a document, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\sigma(q,d) = \sigma'(q,d). \tag{2}$$

Then

$$\underset{q,d}{\mathsf{sim}}(\alpha,\sigma) = \underset{q,d}{\mathsf{sim}}(\alpha,\sigma'). \tag{3}$$

Let us remark that (3) does not entail (2). Diagrams like those in Figure 2 can be helpful to intuitively understand the situation: Figures 2(a) and 2(b) represent the cases in which σ and σ' respectively overestimate and underestimate (or vice-versa) d by the same amount; then the similarity (represented in the figure by the two arcs on the right) of the two systems is the same, but obviously (2) does not hold. Conversely, as stated by the axiom, when (2) holds then (3) holds as well (Figures 2(c) and 2(d)).

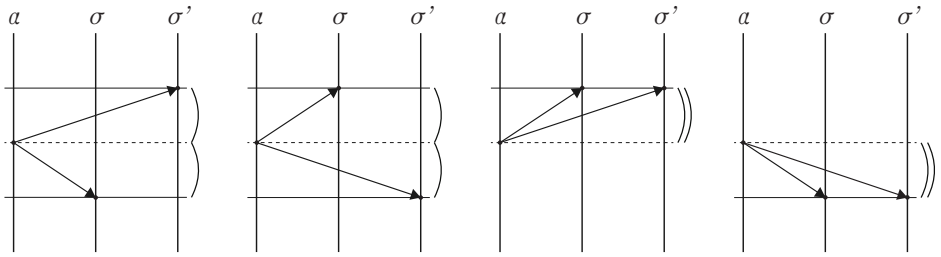
5.2 From Similarity to Metric

5.2.1 *Different systems*

The following axiom sets a constraint on the metric in one of the two last cases of Figure 2 (Figures 2(c) and 2(d)).

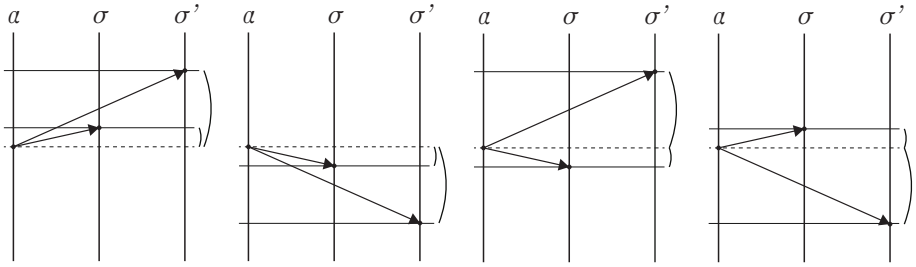
AXIOM 4 (SYSTEMS WITH EQUAL EFFECTIVENESS). *Let q be a query, d a document, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\sigma(q,d) = \sigma'(q,d).$$



(a) Divergent (b) Divergent (c) Overestimate (d) Underestimate

Figure 2: Two systems having equal similarity to α



(a) Overestimated (b) Underestimated (c) Divergent (d) Divergent

Figure 3: Two systems with different similarity to α

Then

$$\underset{q,d}{\mathsf{metric}}\,(\alpha,\sigma) = \underset{q,d}{\mathsf{metric}}\,(\alpha,\sigma').$$

REMARK 1. *Note that by using, in this axiom, a condition like (2) and not like (3) the first two cases of Figure 2 are ruled out, and indeed in those cases we cannot state any constraint on the metric: a recall-oriented metric would give a higher value to a system overestimating all the documents (retrieving all documents means that recall is 1), whereas a precision-oriented metric would do the opposite.*

AXIOM 5 (SYSTEMS WITH DIFFERENT EFFECTIVENESS). *Let q be a query, d a document, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\underset{q,d}{\mathsf{sim}}\,(\alpha,\sigma) > \underset{q,d}{\mathsf{sim}}\,(\alpha,\sigma') \tag{4}$$

and

$$\underset{q,d}{\mathsf{sim}}\,(\sigma,\sigma') > \underset{q,d}{\mathsf{sim}}\,(\alpha,\sigma'). \tag{5}$$

Then

$$\underset{q,d}{\mathsf{metric}}\,(\alpha,\sigma) > \underset{q,d}{\mathsf{metric}}\,(\alpha,\sigma').$$

REMARK 2. *Condition (4) means that σ is less wrong than σ'. The combination of (4) and (5) means that the two systems are wrong in the same direction: if σ overestimates (underestimates) d, then σ' overestimates (underestimates) it even more. Figures 3(a) and 3(b) show these two cases. Condition (4) rules out the other two situations, shown in Figures 3(c) and 3(d), in which no constraint on the metric can be stated for the same reasons mentioned in Remark 1.*

5.2.2 *Different documents*

We now turn to compare a system measurement for two documents d and d'. Let us assume, without loss of generality, that d is more relevant than d' ($\alpha(d) > \alpha(d')$). We can consider two cases:

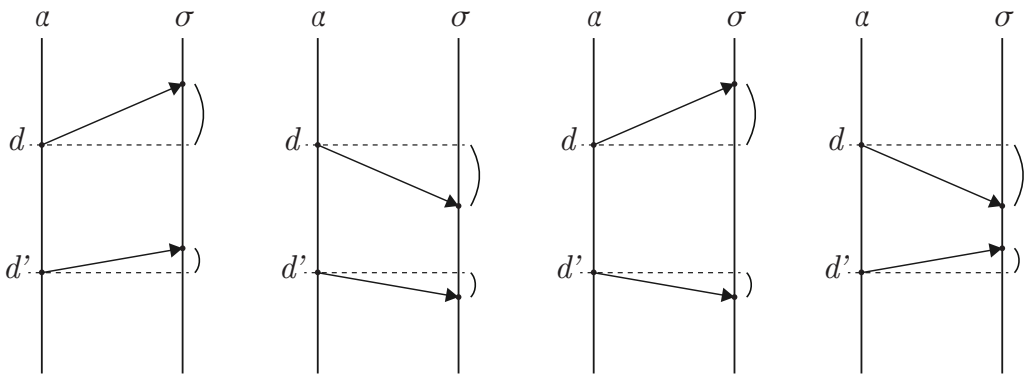
- $\mathsf{sim}_{q,d}\,(\alpha,\sigma) > \mathsf{sim}_{q,d'}\,(\alpha,\sigma)$ (see Figure 4(a));
- $\mathsf{sim}_{q,d}\,(\alpha,\sigma) < \mathsf{sim}_{q,d'}\,(\alpha,\sigma)$ (Figure 4(b)).

In the first case we have a smaller error in the more relevant document and a larger error in less relevant document. In such a case, no constraint can be stated on the metric since, as it is often stated, earlier rank positions are more important than later ones. Conversely, the second case allows to state some axioms. We analyze it and we start by observing that, since the system could overestimate or underestimate the documents d and d', the case of Figure 4(b) can be subdivided into four cases:

- σ overestimates both d and d' (see Figure 5(a));
- σ underestimates both d and d' (Figure 5(b));
- σ overestimates d and underestimates d' (Figure 5(c));
- σ underestimates d and overestimates d' (Figure 5(d)).

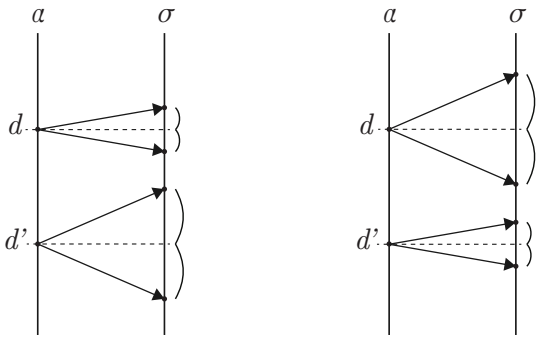
Only the first two cases allow to express some constraints on the metric, again for the same reason of Remark 1 (and 2). We analyze the first two cases. Let us start by noting that: if σ overestimates d then

$$\mathsf{sim}\,(\alpha(d'),\sigma(d)) < \mathsf{sim}\,(\alpha(d),\alpha(d')); \tag{6}$$



(a) Both overestimated (b) Both underestimated (c) Divergent (d) Divergent

Figure 5: Four possible cases



(a) Best similarity top (b) Best similarity bottom

Figure 4: Two documents with different similarity

if σ underestimates d then

$$\mathsf{sim}\,(\alpha(d'),\sigma(d)) > \mathsf{sim}\,(\alpha(d),\alpha(d')); \tag{7}$$

if σ overestimates d' then

$$\mathsf{sim}\,(\alpha(d),\sigma(d')) > \mathsf{sim}\,(\alpha(d),\alpha(d')); \tag{8}$$

and if σ underestimates d' then

$$\mathsf{sim}\,(\alpha(d),\sigma(d')) < \mathsf{sim}\,(\alpha(d),\alpha(d')). \tag{9}$$

We can now state the following two axioms. The first concerns the case of Figure 5(a).

Axiom 6 (Overestimated documents). *Let q be a query, d and d' two document, α a human relevance measurement and σ a system relevance measurements such that*

$$\alpha(d) > \alpha(d'),$$

$$\underset{q,d}{\mathsf{sim}}\,(\alpha,\sigma) < \underset{q,d'}{\mathsf{sim}}\,(\alpha,\sigma)$$

and (6) and (8) hold (i.e., both d and d' are overestimated), then

$$\underset{q,d'}{\mathsf{metric}}\,(\alpha,\sigma) > \underset{q,d}{\mathsf{metric}}\,(\alpha,\sigma).$$

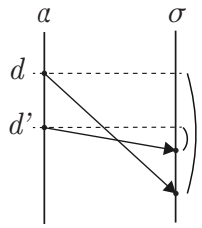


Figure 6: Critical situation

The second axiom concerns the case of Figure 5(b).

Axiom 7 (Underestimated documents). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurements such that*

$$\alpha(d) > \alpha(d'),$$

$$\sigma(d) > \sigma(d'), \tag{10}$$

$$\underset{q,d}{\mathsf{sim}}\,(\alpha,\sigma) < \underset{q,d'}{\mathsf{sim}}\,(\alpha,\sigma),$$

and (7) and (9) hold (i.e., both d and d' are underestimated), then

$$\underset{q,d'}{\mathsf{metric}}\,(\alpha,\sigma) > \underset{q,d}{\mathsf{metric}}\,(\alpha,\sigma).$$

Remark 3. *The condition (10) rules out the critical case in which both documents d and d' are underestimated but there is a "swap" as shown in Figure 6. In such a case, although the similarity is higher for d', no constraint can be imposed on the metric, again for the reason of Remark 1 about top rank positions. In Axiom 6, this additional condition is not necessary, because if both documents are overestimated and similarity is higher for the less relevant document, then no swap is possible.*

In the following we will need to write that a metric value is more affected by a document d than by another document d'. Formally, we define:

Definition 1. *We write that*

$$d \sqsupset_{metric(\alpha,\sigma)} d'$$

if and only if

$$|\underset{q,D\cup\{d\}}{\mathsf{metric}}(\alpha,\sigma)-\underset{q,D}{\mathsf{metric}}(\alpha,\sigma)|>$$
$$|\underset{q,D\cup\{d'\}}{\mathsf{metric}}(\alpha,\sigma)-\underset{q,D}{\mathsf{metric}}(\alpha,\sigma)|$$

(to be read as d affects metric value more than d').

Analogously, we will write $d \sqsupseteq_{metric(\alpha,\sigma)} d'$ and we will also use $\sqsubset$, $\sqsubseteq$, and $\equiv$ with similar meanings. A similar notation holds for queries.

Axiom 8 (System relevance). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurement such that* $\mathsf{sim}_{q,d}(\alpha,\sigma) = \mathsf{sim}_{q,d'}(\alpha,\sigma)$, $\sigma(d) > \sigma(d')$, *and*

$$\alpha(d) \geq \alpha(d'). \tag{11}$$

Then

$$d \sqsupset_{metric(\alpha,\sigma)} d'.$$

This means that if system relevance measures on two documents d and d' are equally correct, and system relevance of d is higher than system relevance of d', then the effectiveness metric should be more affected by d than by d' (provided that d' is not less relevant than d). As already mentioned, it is usually stated that early rank positions affect a metric value more than later rank positions. This can be derived as a corollary of the previous axiom (that states a more general principle, independent of the scales) simply by taking $\mathsf{scale}(\sigma) = [\![Rank]\!]$.

A symmetric axiom can also be stated on user relevance measurement: a metric should weigh more, and be more affected, by more relevant documents. This is perhaps less intuitive than the previous one, but it does indeed seem natural in this framework. Moreover, it is quite easy for an IRS to evaluate a non-relevant document as non-relevant, since the vast majority of documents in the database are non-relevant. Thus, an IRS stating that a non-relevant document is non-relevant is somehow doing an "easy job", and should not be rewarded too much for it. On the other hand it should be rewarded when correctly identifying a relevant document. This is generalized and formalized as follows.

Axiom 9 (User relevance). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurement such that:* $\mathsf{sim}_{q,d}(\alpha,\sigma) = \mathsf{sim}_{q,d'}(\alpha,\sigma)$, $\alpha(d) > \alpha(d')$, *and*

$$\sigma(d) \geq \sigma(d'). \tag{12}$$

Then

$$d \sqsupset_{metric(\alpha,\sigma)} d'.$$

Remark 4. *Conditions (11) in Axiom 8 and (12) in Axiom 9 and are needed to rule out the case in which the two axioms would result inconsistent.*

Finally, the following axiom deals with the last case.

Axiom 10 (Same relevance). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurement such that* $\mathsf{sim}_{q,d}(\alpha,\sigma) = \mathsf{sim}_{q,d'}(\alpha,\sigma)$. *If $\sigma(d) = \sigma(d')$ and $\alpha(d) = \alpha(d')$ then*

$$d \equiv_{metric(\alpha,\sigma)} d'.$$

5.3 Metrics

We now turn to the last set of axioms, that are specifically about metrics.

The following axiom formalizes Swets's properties (see Section 2). To simplify its formulation we denote by $\perp$ the theoretically worst performance, i.e., the relevance measure that gives the worst possible performance according to a given assessor relevance measure.

Axiom 11 (Zero and maximum). *An effectiveness metric should have a true zero in 0 and a maximum value M. The theoretically worst (best) performances $\perp$ should give 0 (M) as the metric value. As a normalization convention let $M = 1$ such that $\forall$ metric, range(metric) $= [0,1]$,* $\mathsf{metric}(\alpha,\alpha) = 1$, *and* $\mathsf{metric}(\alpha,\perp) = 0$.

Axiom 12 (Document monotonicity). *Let q be a query, D and D' two sets of documents such that $D \cap D' = \varnothing$, α a human relevance measurement and σ and σ' two system relevance measurements such that:*[1]

$$\underset{q,D}{\mathsf{metric}}(\alpha,\sigma) \underset{(>)}{\overset{(=)}{\geq}} \underset{q,D}{\mathsf{metric}}(\alpha,\sigma') \tag{13}$$

and

$$\underset{q,D'}{\mathsf{metric}}(\alpha,\sigma) \underset{(=)}{\overset{(=)}{\geq}} \underset{q,D'}{\mathsf{metric}}(\alpha,\sigma'). \tag{14}$$

Then

$$\underset{q,D\cup D'}{\mathsf{metric}}(\alpha,\sigma) \underset{(>)}{\overset{(=)}{\geq}} \underset{q,D\cup D'}{\mathsf{metric}}(\alpha,\sigma'). \tag{15}$$

A similar axiom holds for queries, as follows.

Axiom 13 (Query monotonicity). *Let Q and Q' be two query sets such that $Q \cap Q' = \varnothing$, D a document set, α a human relevance measurement and σ and σ' two system relevance measurements such that:*

$$\underset{Q,D}{\mathsf{metric}}(\alpha,\sigma) \underset{(>)}{\overset{(=)}{\geq}} \underset{Q,D}{\mathsf{metric}}(\alpha,\sigma')$$

and

$$\underset{Q',D}{\mathsf{metric}}(\alpha,\sigma) \underset{(=)}{\overset{(=)}{\geq}} \underset{Q',D}{\mathsf{metric}}(\alpha,\sigma').$$

Then

$$\underset{Q\cup Q',D}{\mathsf{metric}}(\alpha,\sigma) \underset{(>)}{\overset{(=)}{\geq}} \underset{Q\cup Q',D}{\mathsf{metric}}(\alpha,\sigma').$$

These two last axioms can also be interpreted as constraints on the avgD and avgQ functions, respectively.

6. THEOREMS

In this section we demonstrate that the axioms can indeed be used to derive further properties as theorems. For space limitations, we show only a few theorems in this section, and we sketch only one proof; however, the general idea should be clear. We also omit several corollaries that, as already hinted above, can be derived for specific scales.

[1] In this axiom the equal $=$ and less than $<$ signs have obviously to be paired in the appropriate way, "row by row". We use this notation for the sake of brevity and to avoid to state three different and very similar axioms.

THEOREM 1 (UNBALANCED DOCUMENT). *Let q be a query, D a document set, $d \notin D$ a document, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\underset{q,D}{\mathsf{metric}}(\alpha,\sigma) > \underset{q,D}{\mathsf{metric}}(\alpha,\sigma') \tag{16}$$

and

$$\underset{q,D\cup\{d\}}{\mathsf{metric}}(\alpha,\sigma) \leq \underset{q,D\cup\{d\}}{\mathsf{metric}}(\alpha,\sigma') \tag{17}$$

(i.e., σ is more effective, according to the metric, than σ' on D and the situation is reversed on $D \cup \{d\}$). Then

$$\underset{q,d}{\mathsf{metric}}(\alpha,\sigma) < \underset{q,d}{\mathsf{metric}}(\alpha,\sigma') \tag{18}$$

(i.e., σ has to be less effective on q as well).

PROOF. *Let us start by noting that (16) corresponds to (the first or third row of) (13). Now, by way of contradiction let us assume that the conclusion (18) does not hold, i.e.,* $\mathsf{metric}_{q,d}(\alpha,\sigma) \geq \mathsf{metric}_{q,d}(\alpha,\sigma')$. *This corresponds to (14), with $D' = \{d\}$, and with either $>$ or $=$. In both cases, Axiom 12 entails that* $\mathsf{metric}_{q,D\cup\{d\}}(\alpha,\sigma) > \mathsf{metric}_{q,D\cup\{d\}}(\alpha,\sigma')$ *((15) with $D' = \{d\}$), which contradicts (17).* □

A similar theorem, with similar proof (omitted), holds for queries, as follows.

THEOREM 2 (UNBALANCED QUERY). *Let Q be a query set, $q \notin Q$ a query, D a document set, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\underset{Q,D}{\mathsf{metric}}(\alpha,\sigma) > \underset{Q,D}{\mathsf{metric}}(\alpha,\sigma')$$

and

$$\underset{Q\cup\{q\},D}{\mathsf{metric}}(\alpha,\sigma) \leq \underset{Q\cup\{q\},D}{\mathsf{metric}}(\alpha,\sigma').$$

Then $\mathsf{metric}_{q,D}(\alpha,\sigma) < \mathsf{metric}_{q,D}(\alpha,\sigma')$.

THEOREM 3 (CONSISTENT SUBDOCUMENT SET). *Let Q be a query set, D a document set, α a human relevance measurement and σ and σ' two system relevance measurements such that* $\mathsf{metric}_{Q,D}(\alpha,\sigma) > \mathsf{metric}_{Q,D}(\alpha,\sigma')$. *Then*

$$\exists S \subset D \mid \underset{Q,S}{\mathsf{metric}}(\alpha,\sigma) > \underset{Q,S}{\mathsf{metric}}(\alpha,\sigma')$$

(i.e., if σ is more effective than σ' on D, it has to be more effective on a subset of D as well).

REMARK 5. *Recursively applying this theorem we can derive that there is always at least one document in D that is consistent with D. A similar theorem holds for query sets as well.*

THEOREM 4. *Let Q be a query set, D a document set, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\forall q \in Q, d \in D, \underset{q,d}{\mathsf{metric}}(\alpha,\sigma) > \underset{q,d}{\mathsf{metric}}(\alpha,\sigma').$$

Then $\mathsf{metric}_{Q,D}(\alpha,\sigma) > \mathsf{metric}_{Q,D}(\alpha,\sigma')$.

THEOREM 5 (MONOTONICITY OF DOCUMENTS SUBSETS). *Let q be a query, D a document set, S a subset of D, α a human relevance measurement and σ a system relevance measurement such that*

$$\forall d \in S, d' \in D\backslash S, \quad \underset{q,d}{\mathsf{metric}}(\alpha,\sigma) \underset{(\geq)}{>} \underset{q,d'}{\mathsf{metric}}(\alpha,\sigma).$$

Then

$$\forall d' \in D\backslash S, \quad \underset{q,S}{\mathsf{metric}}(\alpha,\sigma) \underset{(\geq)}{>} \underset{q,S\cup\{d'\}}{\mathsf{metric}}(\alpha,\sigma).$$

REMARK 6. *A similar theorem can be stated on queries.*

7. CONCLUSIONS AND FUTURE WORK

Building on the framework based on measure, measurement, and similarity proposed in [6], we have defined some common metrics, proposed some axioms and derived some theorems on IR effectiveness metrics. More generally, when read together with [6], the contribution of this paper is fivefold: (i) the proposal of using measurement to model in a uniform way both system output and human relevance assessment, and the analysis of the different measurement scales used in IR; (ii) the notions of similarity among different measurement scales and the consequent definition of metric; (iii) the definitions of some metrics within the framework; (iv) the axioms; and (v) the theorems.

Several future developments can be imagined, as already listed in [6]. For example, different measurement scales have been proposed in the literature and might be used; axioms for diversity, novelty, and session metrics could be added, and taken into account for specific tasks; and so on.

It is also possible that different sets of axioms can be identified. In this paper we have shown a possible set, and indeed the axioms in [6] are completely different from those proposed here; we leave as future work a detailed comparison of the two sets. The question is left open whether there exist other different sets, and if they are equivalent or perhaps even contradictory, but we remark that the combination of this paper and [6] demonstrates that the framework is expressive and allows to formally reason on effectiveness metrics.

Acknowledgments

We thank Julio Gonzalo and Enrique Amigó for long and interesting discussions, Evangelos Kanoulas and Enrique Alfonseca for helping to frame the Axiometrics research project, Arjen de Vries for suggesting the name "Axiometrics", and organizers of (and participants to) SWIRL 2012. This work has been partially supported by a Google Research Award.

8. REFERENCES

[1] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[2] Enrique Amigó, Julio Gonzalo, and Stefano Mizzaro. A general account of effectiveness metrics for information tasks: retrieval, filtering, and clustering. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1289–1289. ACM, 2014.

[3] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A comparison of evaluation metrics for document filtering. In *CLEF*, volume 6941 of *LNCS*, pages 38–49. Springer, 2011.

[4] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A general evaluation measure for document organization tasks. In Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai, editors, *SIGIR*, pages 643–652. ACM, 2013.

[5] P. Bollmann. Two axioms for evaluation measures in information retrieval. In *SIGIR '84*, pages 233–245, Swinton, UK, 1984. British Computer Society.

[6] Luca Busin and Stefano Mizzaro. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *ICTIR 2013 — Proceedings of the 4th International Conference on the Theory of Information Retrieval*, 2013.

[7] V. Della Mea and S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, 55(6):530–543, 2004.

[8] G. Demartini and S. Mizzaro. A Classification of IR Effectiveness Metrics. In *ECIR 2006*, volume 3936 of *LNCS*, pages 488–491, 2006.

[9] Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In *SIGIR '04*, pages 49–56, New York, NY, USA, 2004. ACM.

[10] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05*, pages 480–487, 2005.

[11] Alistair Moffat. Seven numeric properties of effectiveness metrics. In *AIRS'13*, pages 1–12, 2013.

[12] Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103 (2684):677–80, 1946.

[13] J. A. Swets. Information retrieval systems. *Science*, 141:245–250, 1963.

[14] Amos Tversky. Features of similarity. *Psychological Review*, 84(4), 1977.

[15] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

[16] Wikipedia. Measurement — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Measurement`, 2012. [Last visit: August 2013].

[17] Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.