Axiometrics: Axioms of Information Retrieval Effectiveness Metrics

Eddy Maddalena **Stefano Mizzaro**

Department of Maths and Computer Science
University of Udine
Udine, Italy
Email: eddy.maddalena@uniud.it, mizzaro@uniud.it

Abstract

The evaluation of retrieval effectiveness has played and is playing a central role in Information Retrieval (IR). A specific issue is that there are literally dozens (most likely more than one hundred) IR effectiveness metrics, and counting.

In this paper we propose an axiomatic approach to IR effectiveness metrics. We build on the notions of measure, measurement, and similarity; they allow us to provide a general definition of IR effectiveness metric. On this basis, we provide a definition of some common metrics and we then propose and justify some axioms that every effectiveness metric should satisfy. We also discuss some future developments.

1 Introduction

Effectiveness evaluation is of paramount importance in Information Retrieval (IR). IR has become one of the most evaluation-oriented fields in computer science since the first IR systems (IRS) were developed in the late 1950's. Several effectiveness metrics have been proposed so far. A survey in 2006 (Demartini & Mizzaro 2006) counted more than 50 metrics, taking into account only the system oriented effectiveness metrics. In an extended version of the survey (Demartini et al. n.d.), yet unpublished, about one hundred metrics are collected, let alone user-oriented ones or metrics for tasks somehow related to IR, like filtering, clustering, recommendation, summarization, etc.

As stated for example in (Robertson 2006), there is nothing close to agreement on a common metric that everyone will use. It is a diffuse opinion that different metrics evaluate different aspects of retrieval behavior (Buckley & Voorhees 2000, Robertson 2006). Each of these metrics has its own advantages but also limitations. Metric choice is neither a simple task, nor it is without consequences: an inadequate metric might mean to waste research efforts improving systems toward a wrong target. However, some researchers simply do not investigate into the suitability of the metric for the problem itself and they seem to choose just the most popular metrics for their experiments. We cannot exclude the temptation for researchers to choose, among all available metrics, those that help corroborating their claims, or even to design a new metric to this aim. It is not clear what to do when two metrics disagree.

It is clear that a better understanding of the formal properties of effectiveness metrics would help to avoid wasting time in tuning retrieval systems according to effectiveness metrics inadequate to specific purposes, and it will also induce researchers to make explicit and clarify the assumptions behind metrics. This paper proposes an axiomatic approach to effectiveness metrics, presenting some basic axioms that any reasonable metric should satisfy and that are formulated in a general way.

The paper is structured as follows. Although there has not been much work on formal accounts of IR effectiveness metrics, in Sect. 2 we briefly recall such previous work. To have a common language to state the axioms, we then define a general framework. In Sect. 3 we rely upon the notions of measurement, measure, and scale of measurement to provide a common definition of both the output of an IR system and a relevance assessment by a human judge; the notion of similarity between the two is then analyzed in Sect. 4. An analysis of the measurement scales used in IR and their implications for similarity is also proposed. Measure and similarity allow us to define the notion of effectiveness metric in Sect. 5. In Sect. 6 some metrics are defined within the framework, to demonstrate its effectiveness power. In Sect. 7, some axioms are stated. Conclusions and future work are presented in Sect. 8.

This study is one of the first steps within a larger project, named *Axiometrics*, that is a mix of the words *axioms* (or axiomatic) and *metric* and stands for *axiomatic approach to IR effectiveness metrics*. Axiometrics is one of the research directions discussed and selected as interesting at the recent SWIRL meeting (http://www.cs.rmit.edu.au/swirl12/). One companion paper has already been published on this topic (Busin & Mizzaro 2013); although the present work is similar in the first part, we propose a rather different set of axioms and several refinements and improvements: since Sects. 6 and 7 are completely rewritten, about 40% of the paper is novel.

2 Related Work

Although formal approaches have high importance in the IR field, they have mainly focussed on the retrieval process rather than on effectiveness metrics themselves (see, e.g., (Fang et al. 2004, Fang & Zhai 2005)). However, some research specific to effectiveness metrics does exist, and it is briefly discussed here.

Early attempts have been made by Swets (Swets 1963), who listed some properties of IR effectiveness metrics, and van Rijsbergen (van Rijsbergen 1979, Ch. 7), who followed an axiomatic approach. In (Bollmann 1984), 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 implication is 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) only. 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 (Yao 1995) 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 (Amigó et al. 2009) focus their formal analysis on evaluation metrics for text clustering algorithms finding 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. Still Amigó et al. in (Amigó et al. 2011) discuss, with a similar approach, a unified comparative view of metrics for document filtering. They obtain that no metric for document filtering can satisfy all desirable properties unless a smoothing process is performed. Finally, in an even more recent work (Amigó et al. 2013), they start from a set of formal constraints to define a general metric for document organization tasks, that include retrieval, clustering, and filtering.

3 Measurement

3.1 Measurement, Measures, and Scales

Measurement can be defined as a process aimed at determining a relationship between a physical quantity and a unit of measurement (Wikipedia 2012). In 1946, Stevens (Stevens 1946) defined measurement for social sciences as " the assignment of numerals to objects or events according to some rule". A more recent and widely adopted definition of measurement, proposed by Michell (Michell 1997), is "the numerical estimation and expression of the magnitude of one quantity relative to another".

A particularly discussed issue is how the measurement is expressed. Stevens proposed the four standard *measurement scales* (Stevens 1946): Nominal, Ordinal, Interval, Ratio. This classification has become a tradition in various fields and it provides useful insights (Robertson 2006), although it is rather simple, leaves aside some subtleties, and it has been criticized (Velleman & Wilkinson 1993). Indeed, different scales and classification have been introduced through time (Michell 1997). A slightly different and quite common classification includes: Nominal, Ordinal, Interval, Log-Interval, Ratio, Absolute. A further one is made up of ten levels of measurement (Chrisman 1998): (1) Nominal, (2) Graded membership, (3) Ordinal, (4) Interval, (5) Log-Interval, (6) Extensive Ratio, (7) Cyclical Ratio, (8) Derived Ratio, (9) Counts, and (10) Absolute. Also terminology is discussed: for example, Chrisman's position (Chrisman 1998) is that the term "level" should be preferred to "scale" (in this paper we use "scale").

Another discipline that provides a useful background is Software Measurement (Zuse 1997), where a distinction is made between measurement and measure: a *measurement* is the process through which values are assigned to attributes of entities of the real world; a *measure* is the result of that process, so it is the assignment of a value to an entity with the goal of characterizing a specified attribute. In the rest of this paper we refer to measurement and measure with these meanings.

3.2 IR as Relevance Measurement

The evaluation process in IR is based on two quantities: (i) an automated evaluation, by an IR system, of the possible relevance of a document, and (ii) the user's (or assessor's) estimation of the relevance of a document. We propose to use the above concepts to model these two quantities: given a query, a system tries to *measure* the relevance of the documents to the query, for example to rank the documents; given (a description of) an information need, an assessor tries to *measure* the relevance of the documents to the need. We therefore have two kinds of *relevance measurements* (and *measures* as well): one made by a system and referred to in the following as *system relevance measure(ment)*, and one made by a human and referred to in the following as *user / assessor / human relevance measure(ment)*. A notion of measure / measurement common to both quantities (system and assessor) will allow us to define a notion of similarity among them.

In IR, alternative terms to measurement have been and are used, e.g., assessment, judgment, prediction, score, estimate, amount. Some of what follows could be expressed also without reference to measurement; however, our choice provides a good ground and allows us to exploit the measurement machinery. Yet on terminological issues, we also note that since in the IR field it is common to speak of "effectiveness metrics", we stick with this terminology, also because using "metrics" for the effectiveness metrics avoids confusion with the measure(ment) of relevance made by IRSs and humans. However, let us remark that "effectiveness measure" would perhaps be more appropriate because metrics satisfies some peculiar properties that are not required for measures in their general definitions: not all measures are metrics.

Turning to measurement scales, in IR it is common to use two different scales for system and assessor measurements. Moreover, all the items of the traditional scales make sense, as shown by some examples:

- Nominal: categorizing the documents, e.g., as long / short, or on one specific topic among the topics in a set (e.g., Java vs. C++) or from a specific author, etc. A more subtle question is whether relevance classification (into relevant and nonrelevant) is a nominal classification as well: we will come back shortly on this.
- Ordinal: the usual rank of retrieved documents by an IRS, but also graded relevance assessments like in the four level relevance scale $HRPN$ ('Highly relevant', 'Relevant', 'Partially relevant', and 'Non-relevant').
- Interval: IRSs that try to calibrate their Retrieval Status Values (RSV) may use internal interval scales before providing the ranked output.
- Ratio and Absolute. IRSs that try to estimate the amount of relevance in a document, as has been suggested several times (Swets 1963,

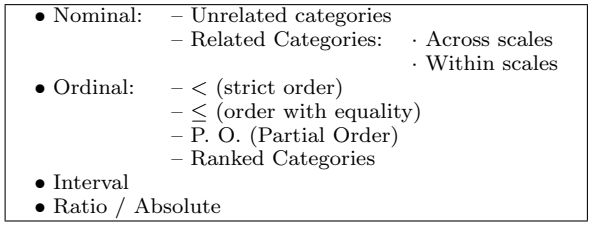


Figure 1: Measurement scales in IR

Della Mea & Mizzaro 2004), or human judges that use magnitude scale estimation (Eisenberg 1988).

On a more careful look, however, there are some peculiar features in IR. The usual relevant / nonrelevant binary scale, when used by both IRS and human assessor, can indeed be seen as a nominal scale (and this leads to defining some well known metrics like precision and recall). However, the situation is more complex. When a graded relevance scale using more than two values is used, it seems unquestionable that it is not nominal but ordinal. This is clear when comparing a graded relevance scale by a human assessor and the usual ranking of documents by an IRS. In other terms, in IR, the categories in a nominal scale are often (although not always) related or, more precisely, ranked. For instance, let us consider a four level relevance scale $HRPN$, that is recently being used and discussed quite often. Within this scale, the four categories are naturally ranked, and thus H is more similar to R, less to P and even less to N (and so on), whereas there is no rank among the categories in a classical nominal scale: the $HRNP$ scale is actually ordinal.

Furthermore, it is quite customary to collapse H, R, P and N into the binary relevance scale R and N as either $HR \rightarrow R$ and $PN \rightarrow N$ or $HRP \rightarrow R$ and $N \rightarrow N$ ("rigid" and "relaxed" mapping in (NTCIR Project 2012)). This means that there exists a relationship across the two scales $HRPN$ and RN, and also that the RN scale is ordinal itself. However, in the classical binary relevance and binary retrieval case, the RN scale is nominal: the scale of a relevance measurement can be nominal or ordinal depending on the scale of the other relevance measurement.

Finally, the rank of categories can be partial: if a document is on Java, a misclassification into the "C++" category is a smaller error than a classification into "Sport", but there is not a total ranking of the scale "C++", "Java", "Sport".

To make things even more complex, the relationship across two scales may concern *different* scales: for example, it is possible to compare a ranked output by an IRS (ordinal scale) with a relevance judgment by a human assessor using magnitude scale estimation (Eisenberg 1988) (ratio or absolute scale).

Thus an IRS and a relevance assessor express measurements of relevance of the documents in a set. Each measurement can be expressed in one of the scales in Fig. 1.

3.3 Notation

In the following, q represents a query, d a document, Q a set of queries, and D a set of documents. We also use subscripts and primes with the obvious meaning. To distinguish different measurements, we denote by ρ a relevance measurement, by σ a system relevance measurement, and by α an assessor, user or, generally, human relevance measurement. We use ρ, α, and σ to represent both the measurement process and the measure itself. Given a query q and a document d, the relevance measurement ρ applied to q and d returns the relevance measure $\rho(q, d)$. To represent the relevance measure of many documents (those in the set D_q) to a single query q we simply write $\rho(q, D_q)$, that can be defined in set theoretic terms as $\rho(q, D_q) = \{\rho(q, d) : d \in D_q\}$, and recursively as $\rho(q, D_q \cup \{d\}) = \{\rho(q, d)\} \cup \rho(q, D_q)$.

The relevance measure of the documents D_q in the set of documents D for the corresponding queries q in Q is $\rho(Q, D)$. It can be defined in set theoretic terms as

$$\rho(Q, D) = \{\rho(q, D_q) : \rho(q, d) \in \rho(q, D_q), q \in Q, \\ D_q \subset D, D_q \text{ is the subset corresponding to } q\},$$

and recursively as $\rho(Q \cup \{q\}, D) = \rho(q, D_q) \cup \rho(Q, D)$.

$\rho(q, d)$ is usually a known value. When it is not known, often it can be compared, for instance by stating that $\rho(q, d) < \rho(q, d')$, i.e., the document d' is more relevant to the query q than the document d, according to the measurement ρ.

The scale of a measurement ρ is denoted by $\mathsf{scale}(\rho)$. Scales are represented by double square brackets $[\![$ and $]\!]$. For example, if ρ is on the above mentioned four level relevance scale $HRPN$, then we write $\mathsf{scale}(\rho) = [\![H, R, P, N]\!]$; a rank scale is denoted by $[\![Rank]\!]$; and so on.

4 Similarity

By exploiting the previous notions we can frame the notion of *similarity* between two relevance measurements. We need a criterion to judge when two relevance measurements, one from the IRS and one from the human assessor, are similar. Ideally, an IRS should use the *same measurement scale* of the human assessor and provide the *same measurement* of the human assessor. However, IRSs are far from being perfect, and therefore the very same measurement is almost never provided. The aim of an IRS is to provide the measurement σ that is most similar to the human assessor / user one α. Moreover, often the scales are different: $\mathsf{scale}(\alpha)$ can be fixed a priori, e.g., when a test collection provides human relevance assessments, and $\mathsf{scale}(\sigma)$ depends on the retrieval algorithm at hand, and different approaches have different scales. Of course, two measurements expressed on two different scales can not be identical (e.g., a rank can not be identical to a measurement expressed on a category scale, the usual ad-hoc retrieval situation). Summarizing, the IRS should provide the measurement that, according to the chosen scales, is the most similar to the human one.

4.1 Notation

The similarity of two relevance measurements ρ (on a set of documents D and a set of queries Q) and ρ' (on D' and Q') is denoted as $\mathsf{sim}(\rho(Q, D), \rho'(Q', D'))$. If $Q = Q'$ then we use the notation $\mathsf{sim}_Q(\rho(D), \rho'(D'))$, or

$$\underset{Q}{\mathsf{sim}}(\rho(D), \rho'(D')),$$

to indicate that Q is common. If $D = D'$ then we write $\mathsf{sim}_D(\rho(Q), \rho'(Q'))$. If both $Q = Q'$ and $D = D'$ we denote the similarity as $\mathsf{sim}_{Q,D}(\rho, \rho')$. We also use the same notation for single documents and queries. For example, if both document and query

sets contain only one element, i.e., $Q = Q' = \{q\}$ and $D = D' = \{d\}$, we write $\mathsf{sim}_{q,d}(\rho, \rho')$. If there is no ambiguity on the sets of queries and documents, then we simply write $\mathsf{sim}(\rho, \rho')$.

Following the notation of Sect. 3.3, the similarity between two systems is given by $\mathsf{sim}(\sigma, \sigma')$, the agreement among judges is $\mathsf{sim}(\alpha, \alpha')$, while the similarity of a human relevance measurement and a system relevance measurement is $\mathsf{sim}(\alpha, \sigma)$, that corresponds to the effectiveness of an IRS in measuring relevance as the user does. In the following, for the sake of simplicity, we mainly deal with this last case, $\mathsf{sim}(\alpha, \sigma)$, but most of the results hold for any pair of relevance measurements.

It is important to note that, in this context, it is not always possible express similarity as a number, i.e., it is possible that we can not know, nor even express, the value of $\mathsf{sim}_{q,d}(\alpha, \sigma)$. More often, similarities can be compared so we can say if $\mathsf{sim}_{q,d}(\alpha, \sigma) < \mathsf{sim}_{q,d}(\alpha, \sigma')$ or $\mathsf{sim}_{q,d}(\alpha, \sigma) > \mathsf{sim}_{q,d}(\alpha, \sigma')$. For example, let us consider two IRSs that behave in exactly the same way on a set of documents D and a set of queries Q, and provide two relevance measurements σ and σ'. Their similarities with an assessor measure α are exactly the same for each document. Therefore, $\mathsf{sim}_{Q,D}(\alpha, \sigma) = \mathsf{sim}_{Q,D}(\alpha, \sigma')$. If a new document $d \notin D$ is available and the two systems have a different behavior on it such that $\mathsf{sim}_{Q,d}(\alpha, \sigma) > \mathsf{sim}_{Q,d}(\alpha, \sigma')$ (σ evaluates the document in a more similar way to the assessor α than σ'), we can now infer that the similarity on the new whole collection of documents $D \cup \{d\}$ is higher for σ than for σ':

$$\underset{Q,D\cup\{d\}}{\mathsf{sim}}(\alpha, \sigma) > \underset{Q,D\cup\{d\}}{\mathsf{sim}}(\alpha, \sigma').$$

Note that we did not need to assign specific values to the various measurements to be able to compare them. Also note that, in general,

$$\underset{q,D\cup\{d\}}{\mathsf{sim}}(\alpha, \sigma) \neq \underset{q,D}{\mathsf{sim}}(\alpha, \sigma) + \underset{q,d}{\mathsf{sim}}(\alpha, \sigma).$$

In the following of this paper we adopt a different approach from (Busin & Mizzaro 2013) 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, since we will be dealing with sets when working on metrics.

4.2 Similarity and Measurement Scale

This generic notion of similarity can be specialized by taking into account the different scales to obtain operational definitions. The scenario is quite complex. Since α is fixed and σ varies (e.g., by choosing a different retrieval algorithm), similarity is not symmetrical: $\mathsf{sim}(\rho, \rho')$ and $\mathsf{sim}(\rho', \rho)$ need to be defined independently. Moreover, each of α and σ can be over nine scales (see Fig. 1): this leads up to $9 \times 9 = 81$ cases, see Tab. 1. Actually, some combinations are ruled out, because either the combination does not make sense (e.g., categories can not be both related and unrelated), or no notion of similarity can be defined (a nominal scale with unrelated categories does not allow to define a notion of similarity). However, all the other combinations make sense: although it might seem strange to have a human assessors that ranks the documents and an IRS that categorizes them into $[\![H, R, P, N]\!]$, this is not impossible in principle. For space limitations, we only hint at how sim can be defined for some pairs of scales; the description is at an intuitive level, but it can be formalized.

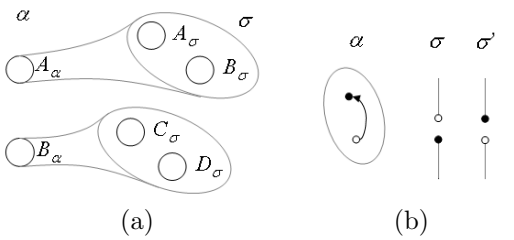


Figure 2: Categories with across-scales relations

If α and σ are expressed on different and unrelated category scales, nothing can be said. For instance, if $\mathsf{scale}(\alpha) = [\![A, B]\!]$ (i.e., α is a category scale with two categories A and B), and $\mathsf{scale}(\sigma) = [\![C, D]\!]$, and there is no relationship neither across scales (i.e., A and B have nothing to do with neither C nor D) neither within scales (i.e., no relations between A and B, nor between C and D), then $\mathsf{sim}(\alpha, \sigma)$ can not be defined. For instance, if σ is a measurement of the topicality of the documents and α classifies the source of the documents (e.g., web page, article, technical report, etc.) then of course α and σ can not be compared.

Even if one and only one of $\mathsf{scale}(\alpha)$ and $\mathsf{scale}(\sigma)$ has some within scale relationship, this is not enough to define any sensible similarity measure (if the two measures share the same categories then a similarity can be defined, but this would mean that an across-scales relationship exists). The following postulate simply rules out the unrelated categories scale from further analysis.

Postulate 1 (Unrelated categories)**.** *Given two measurements, if the scale of at least one of them is an unrelated categories scale, then it is not possible to define a similarity between the two measurements.*

There might exist across-scales relationships: some categories of $\mathsf{scale}(\alpha)$ can be related to some categories of $\mathsf{scale}(\sigma)$, or even equal (i.e., a particular case of across-scales relationships is scales with the same categories). In this case, similarity can be defined and it depends on the number of correctly and wrongly classified documents: moving one document from a wrong category into a correct one (i.e., correctly classifying one more document) increases similarity. For instance, let $\mathsf{scale}(\alpha) = [\![A_\alpha, B_\alpha]\!]$ (i.e., it is expressed on a binary scale), and $\mathsf{scale}(\sigma) = [\![A_\sigma, B_\sigma, C_\sigma, D_\sigma]\!]$. Let us also assume that it makes sense that documents in A_α should also be in A_σ or B_σ, and documents in B_α should also be in C_σ or D_σ (see Fig. 2(a)). In such a case, $\mathsf{sim}(\alpha, \sigma)$ can be defined. Moving items from A_σ to B_σ (or vice-versa) or from C_σ to D_σ (or vice-versa) does not affect $\mathsf{sim}(\alpha, \sigma)$; similarity varies by moving items from A_σ or B_σ to C_σ or D_σ (or vice-versa), or from A_α to B_α (or vice-versa). We can increase $\mathsf{sim}(\alpha, \sigma)$ by modifying σ in σ' in such a way that σ' correctly classifies more documents. Any other modification (e.g., moving a wrongly classified item into another wrong category) does not affect the similarity.

Let us consider another different case. Let σ and σ' be two system relevance measurements such that $\mathsf{scale}(\sigma) = \mathsf{scale}(\sigma') = [\![Rank]\!]$ and let α be a human relevance measurement. If σ' is obtained from σ by swapping two adjacent documents and moving a more relevant document above a less relevant one without affecting anything else (see Fig. 2(b)), then $\mathsf{sim}(\sigma', \alpha) > \mathsf{sim}(\sigma, \alpha)$. The notions of more and less relevant are defined if $\mathsf{scale}(\alpha)$ is any ordinal, interval,

				σ									
				Nominal			Ordinal				Interval	Ratio / Absolute
				Unrelated	Related		$<$	$\leq$	P.O.	Ranked Categ.		
					Acr.	With.						
α	Nominal	Unrelated		•	×	×	•	•	•	•	•	•
		Related	Across	×								
			Within	×		•	•	•	•			
	Ordinal	$<$		•		•						
		$\leq$		•		•						
		P.O.		•		•						
		Ranked Categories		•								
	Interval			•								
	Ratio / Absolute			•								

Table 1: Similarity between relevance measurement scales. Empty cells mean that similarity can be defined. × means that the combination does not make sense. • means that no notion of similarity can be defined.

or ratio / absolute scale. The move of a less relevant document below a more relevant one is equivalent. Other changes can always be obtained by several adjacent swaps (as it is well known, any permutation is a sequence of swaps).

5 Effectiveness Metric

On the basis of the concepts of measurement, measurement scales, and similarity we now turn to modeling the effectiveness metrics itself. An effectiveness metric provides a numerical representation of the similarity between two relevance measurements. A metric is then 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 $\mathbb{R}$):

$$\mathsf{metric} : \alpha \times \sigma \times D \times Q \mapsto \mathbb{R}. \quad (1)$$

A metric is defined 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 $\mathsf{metric}\,(\mathsf{scale}(\alpha), \mathsf{scale}(\sigma), \mathsf{sim}, \mathsf{avgD}, \mathsf{avgQ})$.

In most cases we do not need to specify all the components of the effectiveness metric. Also, we do not need to refer to the metric itself, but rather to the value of the metric in a specific measurement (the context will resolve the ambiguity). For instance we denote with $\mathsf{metric}(\alpha(Q, D), \sigma(Q, D))$ the effectiveness metric value obtained when evaluating σ with respect to α on the set of queries Q and on the set of documents D. Usually, Q and D are common to α and σ, so we often write $\mathsf{metric}_{Q,D}(\alpha, \sigma)$.When not needed, we omit Q and D as in $\mathsf{metric}(\alpha, \sigma)$, and sometimes in place of sets Q and D we also use single elements q and d, as in $\mathsf{metric}_{q,d}(\alpha, \sigma)$.

6 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 2 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 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 (Della Mea & Mizzaro 2004).

7 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 7.1), we then present some axioms that define the relationships between similarity and metrics (Subsection 7.2), and we then present metric-specific axioms (Subsection 7.3).

7.1 Similarity

The first axioms represent basic constraints on similarity, and metrics are not concerned 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*

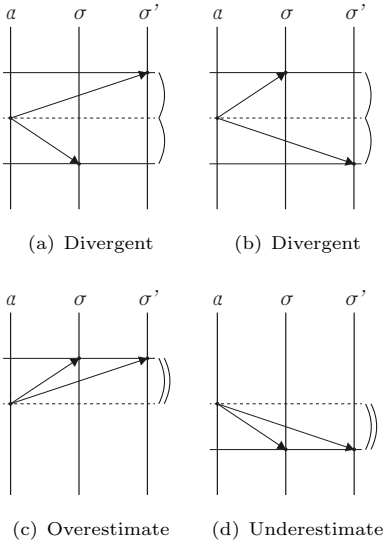
$$\underset{q,d}{\mathsf{sim}}\,(\alpha, \sigma) = \underset{q,d'}{\mathsf{sim}}\,(\alpha, \sigma).$$

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*

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

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

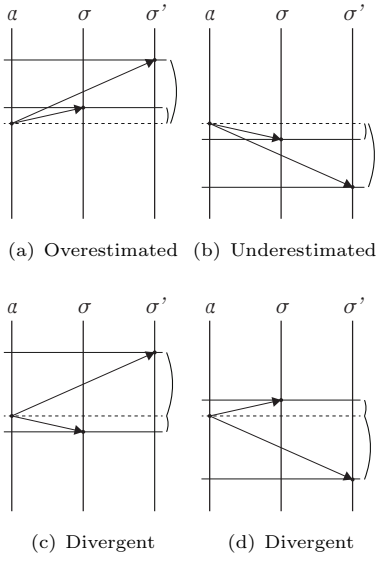
Table 2: Metrics on the basis of their components as per formula. 1



(a) Divergent (b) Divergent

(c) Overestimate (d) Underestimate

Figure 3: Two systems having equal similarity to α



(a) Overestimated (b) Underestimated

(c) Divergent (d) Divergent

Figure 4: Two systems with different similarity to α

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 Fig. 3 can be helpful to intuitively understand the situation: Figs. 3(a) and 3(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 (Figs. 3(c) and 3(d)).

7.2 From Similarity to Metric

7.2.1 Different systems

The following axiom sets a constraint on the metric in one of the two last cases of Fig. 3 (Figs. 3(c) and 3(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*

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

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 Fig. 3 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

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

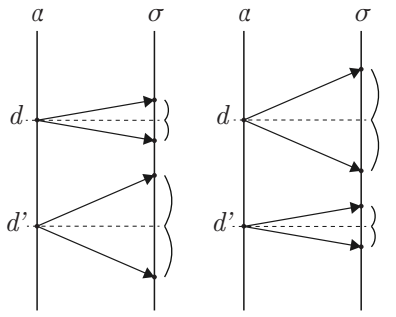
Remark 2. *Condition (4) means that σ is less wrong than σ'. The combination of (4) and (5) means that the two systems are wrong in the same direction: if σ overestimates (underestimates) d, then σ' overestimates (underestimates) it even more. Figs. 4(a) and 4(b) show these two cases. Condition (4) rules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

7.2.2 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:

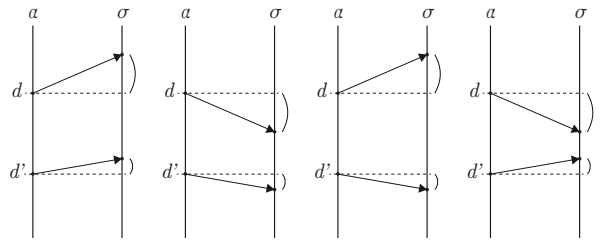
- $\mathsf{sim}_{q,d}\,(\alpha,\sigma) > \mathsf{sim}_{q,d'}\,(\alpha,\sigma)$ (see Fig. 5(a));
- $\mathsf{sim}_{q,d}\,(\alpha,\sigma) < \mathsf{sim}_{q,d'}\,(\alpha,\sigma)$ (Fig. 5(b)).

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 Fig. 5(b) can be subdivided into four cases:



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

Figure 5: Two documents with different similarity



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

Figure 6: Four possible cases

- σ overestimates both d and d' (see Fig. 6(a));
- σ underestimates both d and d' (Fig. 6(b));
- σ overestimates d and underestimates d' (Fig. 6(c));
- σ underestimates d and overestimates d' (Fig. 6(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

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

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 Fig. 6(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

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

The second axiom concerns the case of Fig. 6(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 Fig. 7. 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.*

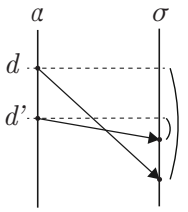


Figure 7: Critical situation

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*

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

7.3 Metrics

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

The following axiom formalizes Swets's properties (see Sect. 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 $\forall\, \mathsf{metric}$, range($\mathsf{metric}$) $= [0,1]$, $\mathsf{metric}(\alpha,\alpha) = 1$, 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]

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

and

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

Then

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

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:*

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

and

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

Then

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

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

8 Conclusions and Future Work

Building on measure, measurement, and similarity, we have defined a framework that has been tested by using it to define some common metrics and to propose some axioms on IR effectiveness metrics. Our contribution is fourfold: (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; and (iv) the axioms themselves.

A future direction concerns the measurement scales: we proposed a set of scales specific for IR, and this proposal has been adequate for this paper; however, the literature on measurement scales is quite rich, and the IR case should perhaps be linked more carefully with it.

An obvious future direction is the definition of some theorems, that we have omitted for space limitations. Again for space limitations we have omitted axioms on diversity, novelty and session metrics, as well as metrics taking into account the notion of difficulty / ease of query and documents. On these issues, something is hinted in (Busin & Mizzaro 2013). Strictness of an effectiveness metric is another interesting property: a metric is strict if having a high value of it implies that also the other metrics will have a high value (i.e., being effective according to a

[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.

strict metric means also being effective according to other metrics).

The set of axioms could be more structured: some of them could be more basic and some others could be derived from the basic ones. It is also possible that more fundamental axioms can be found, for example on the basic notions of measurement and similarity, and that the axioms stated in this paper can indeed be theorems derived from those more fundamental axioms. On similar issues, although we have stated our axioms in a general way, axioms (and theorems) for specific IR tasks and scenarios could probably be derived from the set of general axioms. Of course, it would be interesting to analyze other existing metrics, to see if they satisfy the axioms. It is also possible that different sets of axioms can be identified. In this paper we have shown a possible set, but the question is left open whether there exist other different sets, and if they are equivalent or perhaps even contradictory. Indeed the axioms in (Busin & Mizzaro 2013) are completely different from those proposed here; we leave as future work a detailed comparison of the two sets, but we remark that the combination of this paper and (Busin & Mizzaro 2013) demonstrates that the framework is expressive and allows to formally reason on effectiveness metrics.

Besides axioms and theorems, it would be interesting to think of desiderata (desirable properties, supported by common sense, that could be useful in some scenarios) and empirical properties (those that emerge from data, i.e., from actual test collections and system comparisons). Those could cover aspects like robustness or statistical correlation between effectiveness metrics.

We believe that our research has also shown that basic features of metrics might be quite different from those usually discussed in the classical ad-hoc retrieval situation (binary / category relevance and ranking retrieval), where we usually speak of early rank positions, rank swaps, etc. Finally, it would be interesting to implement a software to analyze metrics by specifying some parameters corresponding to the values of specific components.

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.

References

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

Amigó, E., Gonzalo, J. & Verdejo, F. (2011), A comparison of evaluation metrics for document filtering, *in* 'CLEF', Vol. 6941 of *LNCS*, Springer, pp. 38–49.

Amigó, E., Gonzalo, J. & Verdejo, F. (2013), A general evaluation measure for document organization tasks, *in* G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke & T. Sakai, eds, 'SIGIR', ACM, pp. 643–652.

Bollmann, P. (1984), Two axioms for evaluation measures in information retrieval, *in* 'SIGIR '84', British Computer Society, Swinton, UK, pp. 233–245.

Buckley, C. & Voorhees, E. M. (2000), Evaluating evaluation measure stability, *in* 'SIGIR '00', ACM, New York, NY, USA, pp. 33–40.

Busin, L. & Mizzaro, S. (2013), Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics, *in* 'ICTIR 2013 — Proceedings of the 4th International Conference on the Theory of Information Retrieval'. To appear.

Chrisman, N. R. (1998), 'Rethinking levels of measurement for cartography', *Cartography and Geographic Information Science* **25**(4), 231–242.

Della Mea, V. & Mizzaro, S. (2004), '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.

Demartini, G., Kazai, G. & Mizzaro, S. (n.d.), A survey and classification of information retrieval effectiveness metrics. Draft.

Demartini, G. & Mizzaro, S. (2006), A Classification of IR Effectiveness Metrics, *in* 'ECIR 2006', Vol. 3936 of *LNCS*, pp. 488–491.

Eisenberg, M. B. (1988), 'Measuring relevance judgments', *Informat. Process. & Management* **24**(4), 373–389.

Fang, H., Tao, T. & Zhai, C. (2004), A formal study of information retrieval heuristics, *in* 'SIGIR '04', ACM, New York, NY, USA, pp. 49–56.

Fang, H. & Zhai, C. (2005), An exploration of axiomatic approaches to information retrieval, *in* 'SIGIR '05', pp. 480–487.

Michell, J. (1997), 'Quantitative science and the definition of measurement in psychology', *British Journal of Psychology* **88**(3), 355–383.

NTCIR Project (2012), `http://research.nii.ac.jp/ntcir/index-en.html`. [Last visit: August 2013].

Robertson, S. (2006), On GMAP: and other transformations, *in* 'CIKM '06', New York, USA, pp. 78–83.

Stevens, S. S. (1946), 'On the theory of scales of measurement', *Science* **103 (2684)**, 677–80.

Swets, J. A. (1963), 'Information retrieval systems', *Science* **141**, 245–250.

van Rijsbergen, C. J. (1979), *Information Retrieval*, 2nd edn, Butterworths.

Velleman, P. F. & Wilkinson, L. (1993), 'Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading', *The American Statistician* **47**(1), 65–72.

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

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

Zuse, H. (1997), *A Framework of Software Measurement*, Walter de Gruyter.