Seven Numeric Properties of Effectiveness Metrics

Alistair Moffat

Department of Computing and Information Systems,
The University of Melbourne, Australia

Abstract. Search effectiveness metrics quantify the relevance of the ranked document lists returned by retrieval systems. In this paper we characterize metrics according to seven numeric properties – boundedness, monotonicity, convergence, top-weightedness, localization, completeness, and realizability. We demonstrate that these properties partition the commonly-used evaluation metrics, and hence provide a framework in which the relationships between effectiveness metrics can be better understood, including their relative merits for different applications.

Keywords: Effectiveness metric, precision, NDCG, discounted cumulative gain, rank-biased precision, mean average precision, reciprocal rank.

1 Introduction

Search effectiveness metrics are used to quantify the relevance of the ranked document lists returned by retrieval systems, typically by *scoring* a ranking of length k, where k is a parameter of the experiment and might be 5, or 100, or $1{,}000$, but is unlikely to be the number of documents in the underlying experimental collection. A very large number of metrics have been described in the literature, and used in retrieval experimentation [8]. More are developed each year, including for specialized applications.

An obvious question arises: is the diversity of metrics a consequence of them having different properties and behaviors? Or are they all connected in some fundamental manner? One way of comparing metrics is to look for correlations (and non-correlations) in their numeric scores, arguing that metrics that are non-strongly correlated supply evidence about different system behaviors, while metrics that are correlated are measuring similar aspects, whatever they may be. It is also possible to compare metrics according to their ability to attain similar system orderings on shared retrieval tasks; or based on their likelihood of generating statistically significant pairwise system comparisons; or based on their fit to observed user behavior.

In this paper we take a more fundamental approach, and describe seven simple numeric properties that a metric might or might not have: boundedness, monotonicity, convergence, top-weightedness, localization, completeness, and realizability. Each of the properties is a straightforward attribute; what is surprising is that the commonly-used metrics such as precision, reciprocal rank, average precision, and normalized discounted cumulative gain have different combinations of the seven attributes, and hence have distinctive numerical properties.

R.E. Banchs et al. (Eds.): AIRS 2013, LNCS 8281, pp. 1–12, 2013.

2 Effectiveness Metrics

Preliminaries: We suppose that a document ranking of length k is being scored and has been mapped, perhaps via human judgments, to a real-valued vector $\mathcal{R} = \langle r_i \rangle$. The interpretation is that $0 \le r_i \le 1$ is the utility to the user, in terms of the underlying information need, of the document at depth i in the ranking. It is also assumed that $\mathcal{R}$ can be thresholded in some way to make a vector $\mathcal{B} = \langle b_i \rangle$ of binary relevance values: $b_i = 1$ if $r_i \ge \theta$, and $b_i = 0$ otherwise. Graded relevance assessments provide the judges with multiple options – typically a set of ordered categories – that can be thresholded at different points if binary judgments are required. For example, a graded relevance scale with labels of *None*, *Low*, *Moderate*, *High*, and *Very High* might be translated to the set of utility contributions $\{0.0, 0.1, 0.3, 0.7, 1.0\}$ if numeric relevance scores are required; and might also be thresholded using $r_i \ge$ *Moderate* $\Rightarrow b_i = 1$ in situations in which binary relevance values are desired. Naturally, different mappings (such as the use of $\{0.0, 0.125, 0.25, 0.5, 1.0\}$, or the use of $r_i \ge$ *High*) affect the score that is generated by any particular metric.

In the development that follows, metrics that can only be applied to binary relevance judgments are shown with $\mathcal{B}$ as their argument; metrics that apply to real-valued relevance assessments (including binary-valued ones) are shown with an argument $\mathcal{R}$. Note also that some metrics rely more on k than do others; nevertheless, for consistency all metrics are shown as being evaluated down to a cutoff depth of k. The way in which metrics respond as k is altered is one of the numeric properties discussed in Section 4.

Standard Metrics: *Precision at depth* k is the fraction of the documents in the top k that are relevant (see Büttcher et al. [4] for descriptions of these standard mechanisms):

$$\text{Prec@}k(\mathcal{B}) = \frac{1}{k} \sum_{i=1}^{k} b_i \,.$$

The traditional counterpoint of precision is *recall at depth* k, the fraction of the relevant documents that appear in the first k positions of the ranking, $\text{Recall@}k(\mathcal{B}) = (k/R) \times \text{Prec@}k(\mathcal{B})$, where $R = \sum_{i=1}^{d} b_i$ is the total number of relevant documents in the d-document collection. Using the same definitions, *reciprocal rank at depth* k is

$$\text{RR@}k(\mathcal{B}) = \frac{1}{\min\{i \mid 1 \le i \le k \text{ and } b_i = 1\}} \,;$$

and *average precision at depth* k is given by [2]:

$$\text{AP@}k(\mathcal{B}) = \frac{1}{R} \sum_{i=1}^{k} b_i \times \text{Prec@}i(\mathcal{B}) \,. \tag{1}$$

Note that it is important that the metric evaluation depth k be specified in all cases. Just as Prec@5 is a different metric to Prec@10, so too is AP@5 different to AP@10, and RR@5 different from RR@10. In the computation of AP@k in particular, there are zero contributions assumed from relevant documents outside the top k.

The four metrics introduced so far already display different properties. For example, Recall and AP are not defined if there are no relevant documents, that is, when $b_i = 0$ for all $1 \le i \le d$; and a special case of RR@$k = 0.0$ needs to be defined to cover the same situation. In contrast, provided $k \ge 1$, Prec@k is well-defined even if there are no relevant documents in the top k, or if there are no relevant documents in the whole collection. That is, against just one criterion – whether the metric can be calculated if there are no relevant documents in the collection, the property denoted in Section 4 as *completeness* – there are differences to be found.

User Models: Before proceeding with further metrics, it is useful to note that a *user model* can be associated with each metric [15]. For example, Prec@k corresponds to a user who examines exactly k documents in the ranking, and computes their "expected return" in units of "relevance per document inspected" (abbreviated RPDI for convenience). Similarly, RR@k corresponds to a user who examines at most k documents, and stops either at depth k or as soon as they locate a useful one; the numeric RR score is again an expected return in units of RPDI. Robertson [17] describes a user model in which AP can also be interpreted as an RPDI value, but the user is presumed to have more complex behavior. If that model is adapted to the case of a ranking of depth k, then AP@k corresponds to a user who knows how many relevant documents there are in the collection (the quantity R); chooses at random a number s between 1 and R inclusive; scans the ranking until they have encountered s relevant documents, even if that takes them beyond depth k; and then only actually takes benefit from the relevant documents that occur within the top k. Dupret and Piwowarski extend that model to graded relevance assessments [10].

Discounted Cumulative Gain: The *discounted cumulative gain at depth* k (or DCG@k) metric of Järvelin and Kekäläinen [11] is a variant of precision in which the ranks are top-weighted according to a weighting vector $\mathcal{W}$, so that a relevant document in the top position contributes more to the score than does a relevant document later in the ranking. The corresponding user model assumes that the user views all k documents in the ranking, but places more emphasis on items near the top. The weighting vector $\mathcal{W} = \langle w_i \rangle$ given by Järvelin and Kekäläinen is constant through until depth b, and then decays using logarithms base b; a variant, and the one used in the remainder of this paper, discounts the relevance ranking from the first position, taking $w_i = 1/\log(i+1)$. In this "Microsoft DCG" version, the choice of b is no longer relevant, since logarithms to different bases are related by a multiplicative constant. In all of the numeric examples below, we take $b = 2$. Given a weighting vector $\mathcal{W}$, and a relevance vector $\mathcal{R}$, effectiveness is computed as:

$$\text{DCG@}k(\mathcal{R}) = \mathcal{W} \cdot \mathcal{R}\,,$$

where the $\cdot$ operator represents vector inner product. Note that DCG is the first of the standard metrics that is expressly intended to be used with multi-value relevance assessments rather than binary assessments, and is defined using $\mathcal{R}$ rather than $\mathcal{B}$.

Scaled DCG: A drawback of DCG is that it is unbounded – a DCG effectiveness score might be 3.0, or 23.0, or 123.0. The latter value is attained only when then are 1,000 relevant documents at the head of the ranking, or some even more extensive combination

of relevant documents further down the ranking; nevertheless, it is possible. Indeed, the unbounded nature of the sum $\sum_i 1/\log(i+1)$ means that *any* DCG@k_1 value computed for *any* prefix of length k_1 for one ranking can be exceeded by the DCG@k_2 score for a second ranking that commences with k_1 irrelevant documents, and then contains sufficiently many relevant documents. For example, the DCG@5 score of 1.63 assigned to the ranking $\mathcal{B}$ = "11000" is exceeded by the DCG@11 score assigned to the ranking $\mathcal{B}$ = "00000111111".

One way of introducing a bound is to scale $\mathcal{W}$ according to the sum to k terms of $1/\log(i+1)$, giving rise to a *scaled DCG at depth* k metric:

$$\text{SDCG@}k(\mathcal{R}) = \frac{\text{DCG@}k(\mathcal{R})}{\sum_{i=1}^{k} w_i} = \frac{\text{DCG@}k(\mathcal{R})}{\sum_{i=1}^{k} 1/\log(i+1)}\,.$$

In practical terms SDCG@k has much in common with Prec@k, and can be thought of as being a fixed-depth weighted-precision metric. As an example, $\mathcal{B}$ = "11000" has an SDCG@5 score of $(1.00+0.63)/(1.00+0.63+0.50+0.43+0.39) \approx 0.55$.

More Metrics: If the user will derive full satisfaction if any of the top k documents are relevant, a further metric can be defined: HIT@$k(\mathcal{R}) = \max\{r_i \mid 1 \le i \le k\}$. This metric is appropriate when a single answer is required, such as to a factoid question, or to a named-page finding task.

Following the lead of Järvelin and Kekäläinen [11], Moffat and Zobel [15] introduced *rank-biased precision*. Like SDCG, it is a weighted-precision metric; unlike SDCG, it makes use of an infinite sequence that is convergent, so that there is no requirement for subsequent scaling by a k-dependent denominator:

$$\text{RBP@}k(\mathcal{R}) = \mathcal{W}\cdot\mathcal{R} = (1-p)\left(\sum_{i=1}^{k} r_i p^{i-1}\right)\,.$$

Moffat and Zobel also present a model of user behavior, in which p is the probability that the user will proceed from one document to the next in the ranking, with the RBP score representing the expected return per document inspected, the RPDI units introduced earlier. On any (finite) ranking, the RBP score is a bounded range, which narrows as documents are appended to the ranking. The lower bound of the range is reported as the RBP score, and the difference between it and the upper bound is specified as a *residual* [15]; that is, if we temporarily regard RBP@k as being a real-valued interval, then RBP@$k \supseteq$ RBP@$(k+1)$, and the intervals cannot diverge.

The key distinction between DCG/SDCG and RBP is that the latter uses a weighting vector $\mathcal{W}$ that sums to 1 in the limit, whereas DCG uses a weighting vector that has an unbounded sum, and must be truncated at k terms (hence the definition of SDCG) in order to achieve a bounded sum. There is thus a range of RBP-like metrics, each defined by a convergent infinite sequence of weights. For example, weights of $w_i = 1/(i(i+1))$ define a weighted-precision metric that has similar properties to RBP, since prefix sums of that sequence converge to one. Moffat et al. [14] describe such an inverse-squares weighting function; in the discussion below, we use RBP as a generic label for all metrics derived from infinite decreasing probability distributions.

3 Normalization

AP as a Normalized Metric: Aslam et al. [1] (see also Webber et al. [23]) introduced a *sum of precisions* metric, SP@$k(\mathcal{B}) = \sum_{i=1}^{k} b_i \times \text{Prec@}i(\mathcal{B})$. This computation results in an unbounded metric that shares properties with DCG. It is also related to AP, which is a *normalized* version of SP mapped to the range $[0, 1]$ as a consequence of the division by R. The transformation is useful in one respect, in that 1.0 always represents a "perfect" score; but the division by R means that there is a problem when $R = 0$.

NDCG as a Normalized Metric: Scaled DCG is always in the range $[0, 1]$, and a perfect ranking containing k relevant documents generates an SDCG@k score of 1.0. But if R is smaller than k, SDCG@k cannot be 1.0. An alternative normalization is to divide the actual DCG score by the highest DCG score that could be attained for this particular query, an approach denoted as *normalized discounted cumulative gain at depth* k, or NDCG@k [11]:

$$\text{NDCG@}k(\mathcal{R}) = \frac{\text{DCG@}k(\mathcal{R})}{\text{DCG@}k(\textit{Sort\_Decreasing}(\mathcal{R}))} ,$$

where the denominator represents the DCG@k score that would be attained by an ideal reordering of the ranking, covering all d documents in the collection. Note that NDCG is another metric that is undefined when there are no relevant documents. The "percent perfect" measure of Losee [12] is also a normalized mechanism in this framework.

R-Precision as a Normalized Metric: The same issue also restricts Prec@k: if $R < k$, a score of 1.0 cannot be achieved. Imposing a cap on the score leads to a metric called *R-precision*, the value of Prec@R. For consistency, we also regard RPrec as being evaluated to some depth k, and define RPrec@k to be Prec@k if $k \le R$; to be Prec@R if $k \ge R$; and to be undefined if $R = 0$.

Self Normalization: Computation of AP, NDCG, and RPrec requires that R be known – or, in the case of NDCG and RPrec, that $R \ge k$ be confirmed. In an experimental environment in which a collection of systems are being simultaneously scored, and relevance judgments can be shared across pooled runs, it may indeed be possible to make a reasonable estimate of R [26].

On the other hand, when a small number of systems are being scored, for example, in a simple "before" and "after" experiment, determination of even an approximate value of R might be difficult unless considerably more than k documents are judged for each topic. In such a resource-limited experiment an approach we call *self normalization* is tempting: each topic's relevance ranking is judged to depth k, and then an effectiveness metric is applied over the same k values, using R_k, the number of relevant items present in the top k, instead of R. Then *self normalized DCG at depth* k, or SN-DCG@k, is computed as SDCG@$k(\mathcal{R})$ divided by $\sum_{i=1}^{R_k}(1/\log(i+1))$. For example, SN-DCG@$5$ for the ranking $\mathcal{B} =$ "`10100`" is $(1.0 + 0.50)/(1.0 + 0.63) \approx 0.92$.

Table 1 summarizes the relationship between the three normalized versions of DCG. The sequence of increasingly precise normalizations is intended to adjust the effectiveness score to the bounds imposed by the query and the ranking generated for it. A

Table 1. Variants of DCG. In the case of SN-DCG@k, the scaling denominator is computed based solely on what is returned within the top k documents, and the scaling denominator is oblivious to any relevant documents outside the top k retrieved by the system being evaluated.

Method	Scaling denominator used to adjust DCG@k
DCG@k	No scaling performed
SDCG@k	Max score obtainable by *any* system on *any* query at depth k
NDCG@k	Max score obtainable by *any* system on *this* query at depth k
SN-DCG@k	Max score obtainable by *this* system on *this* query, permuting the top k

similar approach can be used to define *self normalized average precision at depth* k, or SN-AP, where the divisor in Equation 1 is R_k rather than R.

While the SN-DCG approach may appear to be a plausible solution to the question of determining R, it also gives rise to anomalous behavior. Consider the ranking $\mathcal{B}$ = "10101". It has one more relevant document than $\mathcal{B}$ = "10100". But now when SN-DCG is calculated, $R_k = 3$, and so SN-DCG@5 is computed as $1.89/2.13 = 0.88$. That is, the effectiveness score has *decreased*, even though an additional relevant document has *appeared* in the ranking. *Convergence* is one of the seven properties defined in the next section; and is a property that SN-DCG and SN-AP do not have.

4 Numeric Properties of Effectiveness Metrics

Having described a range of metrics, we now enumerate seven properties that might (or might not) be considered desirable in an effectiveness metric. As it turns out, there is tension between these properties, and it is not possible for any metric to attain all of them. Perhaps even more surprising is that the thirteen metrics described in Sections 2 and 3 span a total of ten different combinations of properties (summarized in Table 2).

(1) *Boundedness*: *The set of scores attainable by the metric is bounded, usually in the range* $[0, 1]$. When scores from experiments are being compared, it is desirable for them to be on the same scale. The maximum values of DCG@k and SP@k are functions of R, the number of relevant documents for this query, rather than constant, and so neither of DCG@k and SP@k are bounded. Metrics that are on different numeric scales – and perhaps even those that are not, see Mizzaro [13] – should not have mean scores computed across sets of topics, since they cannot be assumed to have the same units. Other aggregation techniques should be used [16], or standardized versions computed [23].

(2) *Monotonicity*: *If a ranking of length* k *is extended so that* $k + 1$ *elements are included, the score never decreases.* To see why P@k, SDCG@k, and NDCG@k are not monotonic, consider the ranking $\mathcal{B}$ = "11111", which has P@5, SDCG@5, and NDCG@5 scores all of 1.0. But if an additional relevance value is added and the ranking becomes $\mathcal{B}$ = "111110", then the P@6, SDCG@6, and NDCG@6 scores are 0.83, 0.89, and 0.89 (assuming that $R > 5$) respectively, less than the corresponding $k = 5$ scores. On the other hand, RR@6 can never be less than RR@5.

Table 2. Effectiveness metrics, ordered according to their combinations of properties in regard to one possible ordering of properties. All metrics are assumed to be evaluated over a ranking prefix of depth $k \geq 1$, where k is independent of R, the number of relevant documents for the query.

Metric	Bounded	Monoton.	Converg.	Top-wgt.	Localiz.	Complete	Realizb.
DCG@k	No	Yes	Yes	Yes	Yes	Yes	No
SP@k	No	Yes	Yes	Yes	Yes	Yes	No
RPrec@k	Yes	No	No	No	No	No	Yes
SN-DCG@k	Yes	No	No	Yes	Yes	No[d]	Yes
SN-AP@k	Yes	No	No	Yes	Yes	No[d]	Yes
Prec@k	Yes	No	Yes	No	Yes	Yes	No[e]
NDCG@k	Yes	No	Yes	Yes	No	No	Yes
SDCG@k	Yes	No	Yes	Yes	Yes	Yes	No[e]
HIT@k	Yes	Yes	No[b]	No[b]	Yes	Yes	Yes
RR@k	Yes	Yes	No[b]	No[b]	Yes[c]	Yes[c]	Yes
Recall@k	Yes	Yes	Yes	No	No	No	No[f]
AP@k	Yes	Yes	Yes	Yes	No	No	No[f]
RBP@k	Yes	Yes[a]	Yes[a]	Yes	Yes[a]	Yes	No

a. RBP@k yields a constrained range containing the score. The lower end of the range is taken to be the RBP score when a single value is required.
b. RR@k and HIT@k are not convergent or top-weighted because swaps of relevant documents to positions higher up the ranking are not guaranteed to increase the score.
c. RR@k is defined to be zero when there are no relevant documents in the prefix examined.
d. SN-DCG@k and SN-AP@k cannot be calculated when no relevant documents appear in the k-element prefix, even if $R > 0$.
e. SDCG@k and P@k can realize a value of 1.0 only if $k \leq R$.
f. Recall@k and AP@k can realize a value of 1.0 only if $k \geq R$.

Monotonicity is a desirable property when the reported results of an experiment are intended to be a conservative (that is, lower) bound on performance. Use of a monotonic effectiveness metric gives a reader the assurance that, should further relevance judgments be undertaken, the reported scores will increase rather than decrease. Resulted reported using non-monotonic metrics at shallow retrieval depths – for example, NDCG@5 or NDCG@10 – provide little indication as to how the same systems might be assessed in a comprehensive experiment using, say, NDCG@100.

(3) *Convergence*: *If a document outside the top k is swapped with a less relevant one (that is, has a lower r_i value) that is inside the top k, the score strictly increases.* This property complements monotonicity; if a metric is convergent and bounded, scores must strictly converge towards (typically) 1.0 as the density of relevant documents in the top k increases. As is noted in Table 2, there are several non-convergent metrics, most notably the self-normalized variants of NDCG and AP. Both of these can exhibit surprising behavior as relevant documents are inserted into the top k. For example, the ranking

$\mathcal{B}$ = "10000" has a SN-AP@5 score of 1.0, whereas $\mathcal{B}$ = "10001" has a SN-AP@5 score of 0.7. Reciprocal rank is also non-convergent according to this definition: the rankings $\mathcal{B}$ = "01000" and $\mathcal{B}$ = "01100" have the same RR@5 score.

(4) *Top-weightedness*: *If a document within the top k is swapped with a less relevant one (that is, has a lower r_i value) higher in the ranking, the score strictly increases.* A metric is top-weighted if, within the top k, the best score is attained when the relevant documents are in the first positions. The definitions of SDCG@k and RBP@k expressly introduce top-weighting to precision-like metrics, seeking to improve on Prec@k. The "strictly increases" requirement in the definition also implies that RR@k and HIT@k are not top-weighted, since RR@5 on $\mathcal{B}$ = "10001" and $\mathcal{B}$ = "11000" are the same.

Note that all four combinations of convergence and top-weightedness are in evidence in Table 2, and that they are independent concepts. Top-weightedness is similarly independent of monotonicity (as is convergence).

(5) *Localization*: *A score at depth k can be computed based solely on knowledge of the documents that appear in the top k.* The non-localized metrics – RPrec, NDCG, Recall, and AP – typically require specific knowledge of R, the number of relevant documents for the query, or, as a minimum, knowledge that $R \geq k$. As was noted in Section 3, requiring knowledge of R before being able to compute the score means that experimental evaluations are either expensive, with judgments required to depths rather greater than depth k; or must be carried out using approximate values of R derived from pooling; or must be done using self-normalization.

Two of the measures – RBP@k and RR@k – are localized in a slightly specialized sense, in that constrained ranges for the score can be determined after k documents have been judged, even if a single-value score cannot be. In the case of RR@k, either a relevant document is found in the top k, in which case the value of the metric is determined; or if no relevant document is identified, the score is taken to be 0.0. The RBP metric explicitly calculates a range in which the score lies, and as each document is added to the ranking, narrows that range. Regardless of k, the range is always non-zero. When a score value is required, the minimum value in the range is used.

(6) *Completeness*: *A score can be calculated even if the query has no relevant documents.* Metrics that compute normalized scores relative to the best that could be attained for that query – covering Recall, RPrec@k, NDCG@k, AP@k, and the two self-normalized metrics – must of necessity fail to produce a score when $R = 0$. And asserting that when $R = 0$ the score must "of course" be zero is inappropriate, since any ranking at all is "the best that could be attained" if there are no relevant documents, meaning that a score of 1.0 is no less appropriate.[1] The $R = 0$ situation arises in practice in retrieval experimentation, and can be vexing. In the TREC environment, query topics for which no relevant documents are identified in the corresponding collection are removed from the topic set, in order to bypass the awkwardness caused by effectiveness metrics that are not complete. Researchers who work with data subsets must

[1] One might also argue that the empty ranking, containing no documents at all, is more informative than any non-empty ranking of irrelevant documents, and hence should be the only ranking awarded a score of 1.0. As a further option, some researchers might feel that a "divide by zero" error when $R = 0$ is a less uninformative outcome than is a score of zero.

similarly prune topic sets so that they only include queries for which answers are available; one way of rationalizing this need is to argue that a query with no answers cannot differentiate between systems, regardless of what effectiveness score is assigned.

(7) *Realizability*: *Provided that the collection has at least one relevant document, it is possible for the score at depth k to be maximal.* To be realizable, a metric must be capable of generating its maximum value (typically 1.0), even when the number of relevant documents R is larger or smaller than the evaluation depth k. The precision-based metrics Prec@k, SDCG@k, and RBP@k, are unable to always generate a score of 1.0, regardless of how highly the relevant documents are ranked. Nor are Recall and AP realizable, since an evaluation at depth $k < R$ cannot attain a score of 1.0. On the other hand, NDCG generates a score relative to the best that could obtained for this topic, and can yield a score of 1.0 even when there is as few as one relevant document for the topic. Reciprocal rank also falls into this latter category.

All four combinations of completeness and realizability are demonstrated in Table 2, showing that they are independent.

Conflict Between the Properties: Is it possible for a metric to possess all seven of the properties? The answer is no, because monotonicity and convergence between them require that on any ranking that currently contains $k < R$ documents, the score assigned cannot yet be 1.0, making realizability impossible. Hence, of these seven properties, the best that can be achieved is six. Rank-biased precision attains six of the seven, sacrificing realizability. An interesting question is whether there is a metric that retains the other aspects of RBP, but swaps realizability for either monotonicity or convergence; and if such a metric exists, what behavior is implied by the corresponding user model.

5 Subjective Metric Evaluation Criteria

The categorization we have presented is based on objective numeric criteria, and each of the properties is an attribute that a metric either does, or does not, possess. But there are also other subjective criteria that are used when selecting a mechanism (or set of mechanisms) with which to report the results of retrieval experiments. Note that none of the objective criteria summarized in Table 2 implies any of these subjective desiderata.

Meaningfulness: Perhaps the most important attribute of a metric is its plausibility as a measurement tool, that is, whether the scores it generates correlate with the underlying behavior it is intended to represent. If the purpose of the metric is to quantify the overall usefulness of that ranking to the user, then a metric with a user model that doesn't ring true is unlikely to be of interest, regardless of its numeric properties. Normalized metrics, including Recall, have been criticized in this regard – it is difficult to see how the user's perception of the usefulness of a ranking of k documents can depend on the contents of the $d - k$ documents that they are not provided with [25]. More to the point, a range of user studies (see, for example, Turpin and Scholer [22]) have suggested that the link between the effectiveness score of a ranking and its usefulness to users may be tenuous. Even so, the aptness of the metric to the task at hand is an important subjective factor. Web search services are more likely to be measured using HIT@3 or SDCG@5 than via AP@$1{,}000$. A related aspect of this criteria is *scrutability*, whether the score generated by the metric can be readily explained.

Handling Partial Rankings: Another important subjective criterion is the behavior of the metric in the face of incomplete relevance judgments, including the case when a metric is being evaluated to depth k, but the ranking supplied by a system is only of length $j < k$. Buckley and Voorhees [3] introduced the AP-derived BPref metric in order to deal with this problem, with further contributions added by Yilmaz and Aslam [24], and by Sakai [19]. One approach is the use of *condensed rankings*, in which the non-judged documents are removed, and the remaining documents are scored as if that was the list returned by the search system.

A key motivation for RBP was to make explicit, via the provision of a score range, the degree of uncertainty attributable to unjudged documents [15]. Other precision-based metrics such as Prec and SDCG share this ability; whereas computation of score ranges is both more challenging and less informative for metrics such as AP.

Experimental Cost: Researchers designing experiments must construct a judgments budget – an estimate of the cost of carrying out the experiment – as part of their planning. Localized metrics allow such estimates to be made with a degree of confidence not possible with non-localized metrics. Other factors also come in to play when estimating costs, including the fidelity with which results will be presented. Being able to quantify the measurement uncertainty is a useful attribute of weighted-precision metrics; and in the other direction, if a score is required to be known to a given level of uncertainty, that constraint can be used to determine the depth k used in the experimentation [15].

Statistical Properties and Predictivity: Another important facet of metric behavior is the likelihood of statistically significant system differentiations being obtained. All other things being equal, metrics that are predictive of system performance (or system pair relativities) on unseen queries or unseen documents should be preferred to metrics that are not. Several studies have shown AP and NDCG to be useful in this regard [2,18,20,23], assuming that the evaluation depth k (and hence also the pooling depth used to determine R) is sufficiently deepNote that the range of statistical tests that can be used is affected by the metric's numeric properties – not all tests can be applied to bounded values, for example.

Recent Work: Measurement of retrieval effectiveness has been the focus of a range of recent work. For example, the *expected reciprocal rank* metric, ERR [7], is a blend of RR and RBP in which the user is modeled as scanning through to the first relevant document, and then with probability p deciding to scan for the next one, and so on. Other work has sought to compute effectiveness scores based on distributions over parameters (such as the persistence parameter p that governs RBP) so as to better model populations of users [6].

Carterette [5] examines a range of weighted precision metrics, including DCG and RBP, and evaluates them against a set of probability distributions. Carterette concludes that DCG has a number of subjective properties that make it attractive for retrieval experimentation, including that it can be fitted to click log data. Several earlier studies have also made use of click data in order to estimate parameters for user models (and vice versa); see, for example, Dupret and Piwowarski [9].

Most recently, Smucker and Clarke [21] have refined the assumption that user effort can be measured in terms of "documents inspected", and instead suggest that cost

should be based on measured or estimated time. They propose a *time-biased gain* metric which differentiates between long documents and short documents, and between novel documents and repeat occurrences of those documents; Smucker and Clarke estimate parameters for their model through a user study, and via click log analysis.

Moffat et al. [14] consider the relationship between user behavior and effectiveness metrics, arguing that the behaviors modeled by a metric should correspond to the behaviors observed as users carry out search tasks.

6 Discussion

The seven properties that retrieval effectiveness metrics do or do not possess are not completely independent. Nevertheless, the groupings that are apparent in Table 2 show that there are multiple viable combinations. Researchers designing retrieval experiments should thus be alert to the implications associated with the metrics they use, and, conversely, should feel empowered to select metrics that will correctly recognize the behavior that they believe their experiment will reveal. The readers of research papers should be similarly aware of the implications arising from certain choices.

Our primary intention in this work has been to categorize, rather than to criticize. Nevertheless, a caution is in order: the two self-normalized metrics SN-DCG@k and SN-AP@k are counter-intuitive in their behavior, and need to be interpreted with care. Work that reports results using, for example, NDCG@5, should make it clear whether extensive judgments have been performed, or whether SN-DCG@5 is being used. The latter may be less costly to compute, but it is also less well behaved.

As a final remark, note that while the seven numeric criteria are all objective, the determination of them – deciding which properties were important enough to include – has been a subjective exercise. Moreover, the ordering of the columns in Table 2 generates the row ordering shown; with other column orderings yielding different row orderings. Researchers who prefer a different prioritization of the properties (or who feel that some of the listed properties fail to capture meaningful differences between metrics) can reorder the columns (and remove the ones they eschew) in order to focus on the metrics that meet their particular needs. Conversely, there may be additional numeric properties not recognized here – perhaps ones pertinent to metrics not yet considered – that can be added, in order to further refine the categorization.

Acknowledgments. James Allan, Falk Scholer, Paul Thomas, William Webber, and Justin Zobel provided helpful input, as did the anonymous referees via their extensive and thoughtful feedback. This work was supported by the Australian Research Council.

References

1. Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: Proc. SIGIR, Seattle, Washington, pp. 541–548 (2006)
2. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: Proc. SIGIR, Athens, Greece, pp. 33–40 (2000)
3. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proc. SIGIR, Sheffield, England, pp. 25–32 (2004)

4. Büttcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press (2010)
5. Carterette, B.: System effectiveness, user models, and user utility: A conceptual framework for investigation. In: Proc. SIGIR, Beijing, China, pp. 903–912 (2011)
6. Carterette, B., Kanoulas, E., Yilmaz, E.: Simulating simple user behavior for system effectiveness evaluation. In: Proc. CIKM, Glasgow, Scotland, pp. 611–620 (2011)
7. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proc. CIKM, Hong Kong, China, pp. 621–630 (2009)
8. Demartini, G., Mizzaro, S.: A classification of IR effectiveness metrics. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 488–491. Springer, Heidelberg (2006)
9. Dupret, G., Piwowarski, B.: A user browsing model to predict search engine click data from past observations. In: Proc. SIGIR, Singapore, pp. 331–338 (2008)
10. Dupret, G., Piwowarski, B.: A user behavior model for average precision and its generalization to graded judgments. In: Proc. SIGIR, Geneva, Switzerland, pp. 531–538 (2010)
11. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Sys. 20(4), 422–446 (2002)
12. Losee, R.M.: Percent perfect performance (PPP). Inf. Proc. Man. 43(4), 1020–1029 (2007)
13. Mizzaro, S.: The good, the bad, the difficult, and the easy: Something wrong with information retrieval evaluation? In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 642–646. Springer, Heidelberg (2008)
14. Moffat, A., Thomas, P., Scholer, F.: Users versus models: What observation tells us about effectiveness metrics. In: Proc. CIKM, San Francisco, California (to appear, 2013)
15. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Sys. 27(1:2), 1–27 (2008)
16. Robertson, S.: On GMAP: and other transformations. In: Proc. CIKM, Arlington, Virginia, pp. 78–83 (2006)
17. Robertson, S.: A new interpretation of average precision. In: Proc. SIGIR, Singapore, pp. 689–690 (2008)
18. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Inf. Ret. 11(5), 447–470 (2008)
19. Sakai, T.: Alternatives to BPref. In: Proc. SIGIR, Amsterdam, Netherlands, pp. 71–78 (2007)
20. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: Proc. SIGIR, Salvador, Brazil, pp. 162–169 (2005)
21. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Proc. SIGIR, Portland, Oregon, pp. 95–104 (2012)
22. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proc. SIGIR, Seattle, Washington, pp. 11–18 (2006)
23. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: Proc. SIGIR, Singapore, pp. 51–58 (2008)
24. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: Proc. CIKM, Arlington, Virginia, pp. 102–111 (2006)
25. Zobel, J., Moffat, A., Park, L.A.F.: Against recall: Is it persistence, cardinality, density, coverage, or totality? SIGIR Forum 43(1), 3–15 (2009)
26. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proc. SIGIR, Melbourne, Australia, pp. 307–314 (1998)