Computing Maximized Effectiveness Distance for Recall-Based Metrics

Alistair Moffat

Abstract—Given an effectiveness metric $M(\cdot)$, two ordered document rankings X_1 and X_2 generated by a score-based information retrieval activity, and relevance labels in regard to some subset (possibly empty) of the documents appearing in the two rankings, Tan and Clarke's Maximized Effectiveness Distance (MED) computes the greatest difference in metric score that can be achieved that is consistent with all provided information, crystallized via a set of relevance assignments to the unlabeled documents such that $|M(X_1) - M(X_2)|$ is maximized. The closer the maximized effectiveness distance is to zero, the more similar X_1 and X_2 can be considered to be from the point of view of the metric $M(\cdot)$. Here we consider issues that arise when Tan and Clarke's definitions are applied to recall-based metrics, notably normalized discounted cumulative gain (NDCG), and average precision (AP). In particular, we show that MED can be applied to NDCG without requiring an *a priori* assumption in regard to the total number of relevant documents; we also show that making such an assumption leads to different outcomes for both NDCG and average precision (AP) compared to when no such assumption is made.

Index Terms—Information retrieval, effectiveness metric, top-weighted similarity, maximized effectiveness distance, discounted cumulative gain, NDCG, average precision.

1 Introduction

Top-weighted rankings of documents are a common output of information retrieval systems. Given two rankings X_1 and X_2, a natural question to then ask is how "similar" X_1 and X_2 are to each other. Traditional tools such as Kendall's tau cannot be applied to this problem, because the two rankings might not be permutations of each other (documents in either ranking might not appear in the other); because the two lists might be of different or arbitrary length; and because it might be desirable to *top-weight* the comparisons, that is, give more emphasis to differences near the top of the rankings than to differences further down. In response to these three requirements, Webber et al. [1] introduce a computation they call *rank-biased overlap* (RBO), with the property that if $X_1 = X_2$ then RBO is one; if X_1 and X_2 are completely disjoint, then RBO is zero; and if the two lists contain at least some documents in common, but are not equal to each other, then RBO will lie between zero and one.

Tan and Clarke [2] recently observed that in some situations measuring the difference between the sets of elements in the two document rankings via RBO may be less useful than computing the extent to which the two rankings can give rise to differing scores when measured via an effectiveness metric $M(\cdot)$ relative to a set of partial relevance judgments. For example, consider the following two rankings, X_1 and X_2, in which the notation indicates that the top-ranked document in X_1 is A, and that it is non-relevant (that is, has an associated gain of 0); that the second-ranked document in X_1 is B and that its relevance is unknown; and so on:

$$\begin{array}{rcl} X_1 & = & \mathsf{A}_0\ \mathsf{B}_?\ \mathsf{C}_1\ \mathsf{D}_?\ \mathsf{E}_?\cdots \\ X_2 & = & \mathsf{B}_?\ \mathsf{C}_1\ \mathsf{F}_0\ \mathsf{D}_?\ \mathsf{A}_0\cdots \end{array} \tag{1}$$

• *Alistair Moffat is with the School of Computing and Information Systems, The University of Melbourne, Australia. E-mail: See http://people.eng.unimelb.edu.au/ammoffat*

Manuscript received November 29, 2016;revised July 30, 2017.

Note that in this example binary gain values in $\{0,1\}$ are used, but that does not preclude the use of graded relevance categories, with gains that are real-valued. Note also that any documents that appear in both lists must have the same relevance label both times. Suppose further that the metric P@5$(\cdot)$ (the fraction of the first five documents that are relevant) is being used to score the two runs. Based on the known document labels, it is clear that $0.2 \le \text{P@5}(X_1) \le 0.8$, and that $0.2 \le \text{P@5}(X_2) \le 0.6$. To compute the *maximized effectiveness difference* [2] between X_1 and X_2 as measured by P@5, denoted MED(P@5,X_1,X_2), relevance labels are selected for as many of the "free" documents as are required, with the goal of maximizing $|\text{P@5}(X_1) - \text{P@5}(X_2)|$. In the example, the labels associated with documents B and D don't affect that difference, and E is the only document for which a value needs to be bound. Considering the two options, $\mathsf{E} = 0$ and $\mathsf{E} = 1$, the greatest difference is attained when $\mathsf{E} = 1$ and $\text{P@5}(X_1) = \text{P@5}(X_2) + 0.2$, a combination that means that MED(P@5,X_1,X_2) $= 0.2$. In contrast, if the metric is reciprocal rank, RR, then document B plays the (sole) deciding role, and MED(RR,X_1,X_2) $= 0.5$, attained when $\mathsf{B} = 1$.

In a *static weighted-precision metric* the value $M(X)$ over a document ranking $X = \langle x_1, x_2, \ldots, x_i, \ldots\rangle$ is computed as:

$$M(X) = \sum_{i=1}^{\infty} W(i) \cdot g(x_i)\,, \tag{2}$$

where $g(x_i)$ is the real-valued gain accruing in connection with the document x_i at position i of the ranking, usually, but not necessarily, in the range $0 \le g(\cdot) \le 1$, computed via a mapping from categorical relevance grades to numeric gain values; and where $W(\cdot)$ is a fixed weighting function that defines the metric, with $W(i+1) \le W(i)$ and $\sum_{i=1}^{\infty} W(i) = 1$. For example, both precision and rank-biased precision (RBP) [3] fit this structure, but RR does not (it is an *adaptive* weighted-precision metric [4], a category also referred to as *cascade* metrics [5], [6]). Tan and Clarke [2] show that for any static weighted-precision metric $M(\cdot)$,

the value of MED(M, X_1, X_2) can be computed by taking the larger of two differences:

- computing $|M(X_1) - M(X_2)|$ assuming that each uncertain document d is fully relevant ($g(d) = 1$) if and only if d occurs at an *earlier* rank in X_1 than in X_2, and that d is non relevant ($g(d) = 0$) otherwise;
- computing $|M(X_1) - M(X_2)|$ assuming that each uncertain document d is relevant if and only if d occurs at a *later* rank in X_1 than in X_2, and that d is non relevant otherwise.

Uncertain documents that occur at the same rank in both listings need not be considered. For example, referring again to the example in Equation 1, MED(M, X_1, X_2) for a static weighted-precision metric $M(\cdot)$ is computed by considering two alternative document labelings: $\mathsf{B} = 0$, $\mathsf{E} = 1$; and $\mathsf{B} = 1$, $\mathsf{E} = 0$. Once that calculation has been done, the maximum *residual* – the uncertainty arising in metrics such as RBP as a result of the infinite tail beyond the evaluation depth k in the RBP ranking, with weight $\sum_{i=k+1}^{\infty} W(i)$ – needs to be included as a further adjustment [2].

2 Combining MED and NDCG

Järvelin and Kekäläinen [7] introduce a sequence of three effectiveness metrics: cumulative gain, CG; discounted cumulative gain, DCG; and normalized discounted cumulative gain, NDCG. Originally presented as vector-based computations that span every possible evaluation depth k, all three are usually now evaluated to some fixed depth k in the same way as precision is, with the choice of k influencing the behavior of the metric. To compute DCG@k, Equation 2 is applied to a sequence of gain values using the weightings:

$$W(i) = \begin{cases} 1/\log(1+i) & \text{when } i \le k \\ 0 & \text{when } i > k\,. \end{cases} \tag{3}$$

In this computation, $\sum_{i=1}^{\infty} W(i) \approx \ln k > 1$, and hence DCG is not a static weighted-precision metric according to the definition given in Section 1. To achieve a static weighted-precision metric referred to as *scaled discounted cumulative gain* (see, for example, Moffat [8]), or SDCG@k, a slightly different weighting is employed:

$$W(i) = \begin{cases} 1/(S_k \cdot \log(1+i)) & \text{when } i \le k \\ 0 & \text{when } i > k\,, \end{cases} \tag{4}$$

where S_k is a scaling constant, dependent solely on k, to ensure that $\sum_{i=1}^{\infty} W(i) = 1$ and hence that SDCG@k is in $[0,1]$:

$$S_k = \sum_{i=1}^{k} \frac{1}{\log(1+i)}\,. \tag{5}$$

Table 1 lists the scaling factors S_k and weight vectors $W(i)$ for the first few values of k. As k increases, weight is shifted down the ranking and the metric becomes less heavily top-weighted. With this modification, MED(SDCG@$k, \cdot, \cdot$) can be computed by the mechanism of Tan and Clarke [2], summarized in Section 1.

The truncated metric NDCG@k also introduced by Järvelin and Kekäläinen [7] uses *normalization* (rather than scaling) to bring the DCG@k score into the range $[0,1]$. To understand the difference, note that the conversion factor S_k used in SDCG (Equation 4) represents the maximum DCG score that can ever be attained on any ranking of k items, and that an SDCG@k score of 1.0 corresponds to the situation in which every one of the first k documents is maximally relevant. In contrast, NDCG@k employs a conversion factor usually referred to as "ideal DCG"

TABLE 1
Weights $W(i)$ for SDCG@k for different values of k. Where $W(i)$ is unspecified, the weight is zero.

k	S_k	$W(i)$, given k					
		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$
1	1.000	1.000					
2	1.631	0.613	0.387				
3	2.131	0.469	0.296	0.235			
4	2.562	0.390	0.246	0.195	0.168		
5	2.948	0.339	0.214	0.170	0.146	0.131	
6	3.305	0.303	0.191	0.151	0.130	0.117	0.108

that aggregates the *known* relevance information for the *particular* topic in question, in order to provide an effectiveness score range in which 1.0 represents "the best that can be done for *this* ranking", and may not require that all of the first k documents be relevant.

In particular, suppose that the information retrieval topic that led to the two rankings being compared has R relevant documents across the collection (again, we employ binary relevance in our discussion, but non-binary relevance values are readily incorporated). Then NDCG@k is also computed using Equation 4, but with one critical change: rather than use S_k as the conversion factor, $S_{\min\{k,R\}}$ is employed. This distinction between SDCG and NDCG is a critical one, and has consequences for the computation of MED. Specifically NDCG is a *recall-based* effectiveness metric, and does not meet the requirements of the static weighted-precision approach outlined in Section 1, since now $W(i)$ is not a function of k alone – it is also dependent on R, the number of relevant documents available to be found.

In their examination of NDCG, Tan and Clarke [2] choose to normalize by S_k rather than by ideal DCG. That is, they conflate NDCG and SDCG, and assert that SDCG@k is a sufficient proxy for NDCG@k that they can be used interchangeably. Tan and Clarke then note that SDCG@k is a static weighted-precision metric in terms of the definition given in Section 1, and hence can be readily used as the basis for computation of MED scores. Our key claim in the current paper is that Tan and Clarke's conflation of SDCG@k and NDCG@k is neither desirable nor necessary – that MED(SDCG@$k, \cdot, \cdot$) is different to MED(NDCG@$k, \cdot, \cdot$), and that in any case, MED(NDCG@$k, \cdot, \cdot$) can be computed efficiently.

Algorithm 1 supports the second of those two claims, and describes a process for calculating MED(NDCG@k, Y, Z) for input sequences Y and Z of length k. Recognizing that an NDCG score consists of a numerator part and an ideal DCG denominator part, it computes a sequence of "possible" NDCG values by tentatively allowing more and more of the free documents to be assigned relevant labels. The tentative MED score in each configuration is then the difference of two numerators, one for each of Y and Z, divided by the same shared denominator. In Algorithm 1, that difference in numerators is tracked in the variable *numer* as the denominator *denom* increases. The initial value of *denom* is computed from R_J (steps 18–19), based on the set of supplied relevance labels (step 1). The progression of denominator scores followed by *denom* is shown in the column "S_k" in Table 1; each time any unlabeled document is tentatively assigned as relevant, variable *denom* progresses to the next value down that column.

The numerators that match these tentative assignments are determined in a greedy manner, by picking at each iteration the available free document with the highest difference between its numerator weight in Y and its numerator weight in Z, and adding

Algorithm 1 Computing MED(NDCG@k, Y, Z), the greatest difference possible between the NDCG@k scores for Y and Z.

Input: Rankings $Y = \langle y_1, \ldots, y_k \rangle$ and $Z = \langle z_1, \ldots, z_k \rangle$ of length k; and labels J, where $J[d] \in \{0, 1, ?\}$ represents what is known in regard to the gain of document d.

1: $R_J \leftarrow |\{d \in J \mid J[d] = 1\}|$ ▷ the minimum number relevant
2: $F \leftarrow \{d \in Y \cup Z \mid J[d] = ?\}$ ▷ the set of free docs
3: $baseY \leftarrow baseZ = 0$
4: $wY[d] \leftarrow wZ[d] \leftarrow 0$, for all $d \in F$
5: **for** $i \leftarrow 1$ **to** k **do**
6: $wY[y_i] \leftarrow 1/\log_2(1+i)$ ▷ doc weights in Y and Z
7: $wZ[z_i] \leftarrow 1/\log_2(1+i)$
8: **if** $J[y_i] = 1$ **then** ▷ scores for Y and Z
9: $baseY \leftarrow baseY + 1/\log_2(1+i)$
10: **if** $J[z_i] = 1$ **then**
11: $baseZ \leftarrow baseZ + 1/\log_2(1+i)$
12: **for** $d \in F$ **do**
13: $diff[d] \leftarrow wY[d] - wZ[d]$
14: sort F so that $diff[F_i] \geq diff[F_{i+1}]$, most positive to
15: most negative
16: $biggest \leftarrow 0$
17: $numer \leftarrow baseY - baseZ$ ▷ initial numerator
18: $R \leftarrow \min\{k, R_J\}$
19: $denom \leftarrow \sum_{i=1}^{R} 1/\log_2(1+i)$ ▷ initial denominator
20: **if** $denom > 0$ **and** $numer/denom > biggest$ **then**
21: $biggest \leftarrow numer/denom$ ▷ better MED found
22: // consider each of the positive differences that increase
23: // the score of Y relative to Z
24: **for** $d \in F$, in decreasing sorted order of $diff[d]$ **do**
25: **if** $diff[d] \leq 0$ **then**
26: **break** ▷ no more positive differences
27: $numer \leftarrow numer + diff[d]$
28: **if** $R < k$ **then**
29: $R \leftarrow R + 1$ ▷ another relevant
30: $denom \leftarrow denom + 1/\log_2(1+R)$
31: **if** $numer/denom > biggest$ **then**
32: $biggest \leftarrow numer/denom$ ▷ better MED found
33: // now consider the other differences, seeking to increase
34: // the score of Z relative to the score of Y
35: set $diff[d] \leftarrow -diff[d]$ for all $d \in F$, and reverse F,
36: so that $diff[F_i] \geq diff[F_{i+1}]$
37: $numer \leftarrow baseZ - baseY$ ▷ the other initial numerator
38: repeat steps 18 to 32, continuing to maintain *biggest*
39: **return** *biggest*

that difference to *numer*. That is, if another document is to be labeled as being relevant, then it should be the one that creates the greatest positive difference in NDCG numerators, a greedy approach that is correct because addition of the differences results in the same sum, regardless of the order in which the additions take place. In the pseudo-code, steps 5 to 9 compute into $wY[d]$ and $wZ[d]$ the weights of document d in Y and Z respectively, leaving them at zero if that document does not appear in that ranking. Once the document weights have been determined, the difference between those two weights is computed at step 13 as the amount of numerator change associated with making that relevance assignment, and is stored in the array $diff[d]$ as a signed value associated with document d. Steps 5 to 9 also compute the two *base* scores, the numerators for Y and Z arising from the initial document labels provided via the initial judgment set J.

Steps 17 to 20 initialize the process of computing all plausible ratios of $numer/denom$, by setting *biggest* to the difference in the two rankings' initial NDCG values based on J. Labels are then assigned to free documents so as to increase NDCG@$k(Y)$ − NDCG@$k(Z)$. Only positive values of *diff* can increase that value; at step 27 the next largest available *diff* is selected and added to *numer*, with the corresponding increase to *denom* applied at step 30. The largest $numer/denom$ value is tracked throughout; at some point, as yet another document is assigned a notional "relevant" label, *biggest* takes on a maximum value, and is the largest value possible for NDCG@$k(Y)$ − NDCG@$k(Z)$. Note that the denominator in NDCG@$k(\cdot)$ is capped at S_k, even if further free documents are assigned "relevant" labels.

Steps 35 to 38 then repeat the process, but now seeking to maximize NDCG@$k(Z)$ − NDCG@$k(Y)$, by accumulating the originally-negative differences in magnitude order. To do that, the sequence of *diff* values is negated, the sequence of ordered-by-difference document identifiers in F is reversed, and *numer* is initialized to $baseZ - baseY$.

At the end of the two passes through *diff*, each value of which is added at most once, variable *biggest* contains the largest ratio of numerator to denominator that was constructed at any stage, and hence can be returned at step 39 as the required value of MED(NDCG@k, Y, Z). Note that Algorithm 1 also correctly handles the case where the judgments J include "1" labels on documents that do not appear in either of the two rankings; that is why there is a "min" operator at step 18.

Algorithm 1 requires $O(k \log k)$ time, with the sorting process at step 14 the only component that is non-linear in the length k of the two rankings, provided that J, wY, wZ, and *diff*, all of which are indexed by document identifiers d, are implemented in a manner that allows lookup operations to be carried out in constant time – for example, using hashing.

Consider the input rankings X_3 and X_4, of length $k = 10$:

$$\begin{aligned} X_3 &= \mathsf{A}_1\ \mathsf{B}_?\ \mathsf{C}_?\ \mathsf{D}_?\ \mathsf{E}_0\ \mathsf{F}_?\ \mathsf{G}_?\ \mathsf{H}_?\ \mathsf{J}_?\ \mathsf{K}_?\ \cdots \\ X_4 &= \mathsf{A}_1\ \mathsf{D}_?\ \mathsf{B}_?\ \mathsf{E}_0\ \mathsf{C}_?\ \mathsf{G}_?\ \mathsf{F}_?\ \mathsf{J}_?\ \mathsf{L}_?\ \mathsf{H}_?\ \cdots \end{aligned} \tag{6}$$

Table 2 traces the execution of Algorithm 1 for these two sequences. Each row in the table represents one iteration of the main loop at step 24, and shows the document d being considered at that step, the two ranking weights associated with that document, the *diff* value that results (including its original sign before step 35), the *numer* and *denom* values that arise, and then, finally, the ratio $numer/denom$. The largest value of $numer/denom = 0.235$ occurs while the negative differences are being processed, and is the value of MED(NDCG@k, Y, Z). It arises when documents L and D are assigned "1" labels, and hence requires that $R = R_J + 2 = 3$.

Finally in connection with Algorithm 1, note that all gain assignments made during the maximization process are of necessity of the "maximally relevant" category, and that this does not in any way preclude the use of graded relevance values within the pre-existing relevance judgments. That is, provided R_J and R are computed in accordance with the pre-existing relevance grades (steps 1, 18, and 19), and provided a suitable recomputation is carried out to account for one more maximally relevant document (step 30), Algorithm 1 can also be applied when graded NDCG is being computed.

TABLE 2
Computing MED(NDCG@10, X_3, X_4), for the two sequences X_3 and X_4 shown in Equation 6.

d	$wY[d]$	$wZ[d]$	$diff[d]$	R	$numer$	$denom$	$numer/denom$
processing the positive differences							
			initial values	1	0.000	1.000	0.000
K	0.289	0.000	+0.289	2	0.289	1.631	0.177
B	0.631	0.500	+0.131	3	0.420	2.131	0.197
C	0.500	0.387	+0.113	4	0.533	2.562	0.208
H	0.315	0.289	+0.026	5	0.560	2.948	0.190
F	0.356	0.333	+0.023	6	0.582	3.305	0.176
processing the negative differences							
			initial values	1	0.000	1.000	0.000
L	0.000	0.301	–0.301	2	0.301	1.631	0.185
D	0.431	0.631	–0.200	3	0.501	2.131	**0.235**
G	0.333	0.356	–0.023	4	0.524	2.562	0.205
J	0.301	0.315	–0.014	5	0.539	2.948	0.183

3 COMBINING MED AND AP

Average precision (AP) is nominally defined in terms of all documents in the collection (see, for example, page 71 of Büttcher et al. [9]). Given a set of binary gain values $g(\cdot) \in \{0,1\}$ for a ranking X, AP is computed as:

$$\mathrm{AP}(X) = \frac{1}{R}\sum_{i=1}^{N} g(x_i)\cdot \mathrm{P@}i(X)\,, \tag{7}$$

where N is the number of documents in the collection; R is the number of them that are relevant; and where P@i is the metric "precision at depth i". The normalization by R represents a scaling by "what is possible", and means that AP, like NDCG, is a recall-based metric.

Two different approaches have emerged to compute AP@k, that is, an average precision score for a k-element ranking that is a prefix of a whole-of-collection ordering. The first is the approach embodied in the program `trec_eval`[1], where the normalization by R is employed without modification, regardless of the relationship between k and R:

$$\mathrm{AP@}k(X) = \frac{1}{R}\sum_{i=1}^{k} g(x_i)\cdot \mathrm{P@}i(X)\,. \tag{8}$$

If $k < R$, the metric score generated by Equation 8 cannot be 1.0. The other option – see, for example, Sakai [10] – is to normalize by the smaller of R and k:

$$\mathrm{AP@}k(X) = \frac{1}{\min\{k,R\}}\sum_{i=1}^{k} g(x_i)\cdot \mathrm{P@}i(X)\,, \tag{9}$$

thereby ensuring that a k-item ranking in which every document is relevant is assigned a score of 1.0. Lu et al. [11] explore the difference between these two formulations.

Tan and Clarke [2] describe unbounded AP@k in a form equivalent to Equation 7, but then choose to introduce a third "at-k" variant (their Equation 19):

$$\mathrm{AP@}k(X) = \frac{1}{k}\sum_{i=1}^{k} g(x_i)\cdot \mathrm{P@}i(X)\,, \tag{10}$$

where R is ignored in the normalization process, and hence, when $R < k$, metric scores of 1.0 cannot be achieved. Tan and Clarke indicate that this version is required so that the MED transformation can be applied.

1. http://trec.nist.gov/trec_eval/

TABLE 3
Computed score differences for four different metrics at depth $k = 10$ and four different assignments of relevance values to the nine free variables in Equation 6. The AP computation is based on Equation 8.

Relevance assignment										R	$\lvert M(X_3) - M(X_4)\rvert$			
B	C	D	F	G	H	J	K	L		SDCG	NDCG	SSP	AP
1	1	0	1	0	1	0	1	0	6	**0.128**	0.176	0.155	0.259
0	0	1	0	0	0	0	0	1	3	0.110	**0.235**	0.083	0.278
1	1	0	1	1	1	1	1	0	8	0.120	0.138	**0.161**	0.201
1	1	0	0	0	0	0	1	0	4	0.117	0.208	0.113	**0.283**

The relationship between the metric defined by Equation 10 and AP@k as defined by Equation 8 is similar to the relationship between metrics SDCG and NDCG, discussed in Section 2. In both cases the original metric has been modified to remove the normalizing component (replacing ideal DCG by k in the case of NDCG, and replacing R by k in the case of AP), and in both cases, the result is a recall-free computation. Following the terminology of Webber et al. [12] (see also Moffat [8]), we refer to Equation 10 as SSP@k, where SSP stands for *scaled sum of precisions*. It is the scaled form of *sum of precisions*, which, like DCG, is unbounded:

$$\mathrm{SP}(X) = \sum_{i=1}^{N} g(x_i)\cdot \mathrm{P@}i(X)\,. \tag{11}$$

However, SSP is neither recall-based, nor a static weighted-precision metric; and has different characteristics to both AP and to SDCG. That is, computing MED(SSP@$k,\cdot,\cdot$) does not provide guidance as to the maximum difference that can arise in AP scores.

Table 3 provides a numeric example, showing the outcome of four distinct relevance assignments to the free documents in the lists X_3 and X_4 in Equation 6. Each value in the right-hand half of the table is a difference $|M(X_3) - M(X_4)|$, where $M(\cdot)$ is one of SDCG, NDCG, SSP, and AP (as described in Equation 8), with $k = 10$. The bold values are the four MED values for those metrics; in this example, each MED value is a strict maximum, and there are no other relevance assignments that result in scores that are equal to the four bold values. Each of the three other values in each column is strictly less than the listed MED value in that column. Note how the MED transformation allows SDCG and SSP to be less frugal than NDCG and AP in their assignments of relevance to the free documents. That difference is a direct

consequence of them being scaled metrics rather than recall-based normalized ones.

The diversity of "worst" relevance assignments listed in Table 3, the fact that they depend on different assignments of labels to free documents, and the fact that they have different values for R, suggests that no particular information can be gleaned by computing MED(SDCG@$k,\cdot,\cdot$) instead of MED(NDCG@$k,\cdot,\cdot$), nor by computing MED(SSP@$k,\cdot,\cdot$) rather than MED(AP@$k,\cdot,\cdot$). That is, we argue that while MED(SDCG@$k,\cdot,\cdot$) and MED(SSP@$k,\cdot,\cdot$) may have merit in their own right, they should not be regarded as surrogates for MED(NDCG@$k,\cdot,\cdot$) and MED(AP@$k,\cdot,\cdot$).

For rankings with fewer than around 25 to 30 free documents, exhaustive evaluation over all combinations of relevance assignments is viable. If MED(AP@$k,\cdot,\cdot$) as defined by Equations 8 or 9 is required for larger numbers of free variables, the tabu-search approach described by Tan and Clarke [2] for computing MED(SSP@$k,\cdot,\cdot$) can be employed, iterating as appropriate over possible values for R from R_J (see step 1 in Algorithm 1) through to $R_J + 2k$ (at most $2k$ further documents need to be considered for relevance). On the other hand, note that in the case of the untruncated AP (Equation 7), the value of MED(AP,$\cdot,\cdot$) is unlikely to be economically computable, since it requires rankings that span the entire collection, with an evaluation depth k that is very large. Nor is it likely to be especially meaningful.

4 Discussion

Metrics and the computations derived from them evolve in response to a factors arising from their use or application. For example, the NDCG described in Section 2 differs from that described by Järvelin and Kekäläinen [7], in that it removes the original parameter b and uses the "+1" denominator instead. Also worth noting is that Järvelin and Kekäläinen's original 2002 suggestion for an "at k" version of NDCG was to compute (in terms of the definitions given in Section 2) the value $(1/k)\sum_{i=1}^{k}$ NDCG@i, that is, take the mean over all prefixes of the ranking of length up to k; this is a part of Järvelin and Kekäläinen's work that has never been adopted for practical use. Similarly, as was noted in Section 3, there are differing opinions as to how AP@k should be computed, compared to the unbounded version.

Nevertheless, we argue that the distinction between recall-based metrics and weighted-precision metrics is a long-standing and fundamental one, and that the choice between local normalization to obtain a recall-based score (NDCG and AP), and global normalization/scaling to obtain a weighted-precision score (SDCG, SSP, and RBP), is one that is worth retaining as the MED transformation is applied, and should not be put at risk by ambiguous terminology. Hence our desire to clarify that what Tan and Clarke [2] refer to as "MED-NDCG" is, in fact, "MED-SDCG"; and that what they refer to as "MED-AP" is not the maximized difference between AP scores, it is "MED-SSP".

It might be argued that the distinction between NDCG@k or AP@k on the one hand, and SDCG@k or SSP@k on the other, is unimportant in practical situations in which relevance judgments are available, since R will be greater than k for typical values of k such as 10 or 20. In response, Figure 1 plots the judgments associated with the TREC 2013 Web Track, which was based on the ClueWeb12 collection, and used metrics ERR@20 and NDCG@20 as the system evaluation tools [13]. A total of fifty topics were judged in that activity, some to depth 10 and some to depth 20, and graded judgments were created, with "1" indicating "Relevant, possibly minimally"; "2" indicating "Highly relevant"; and "3" and "4" indicating "Key" and "Navigational", respectively.

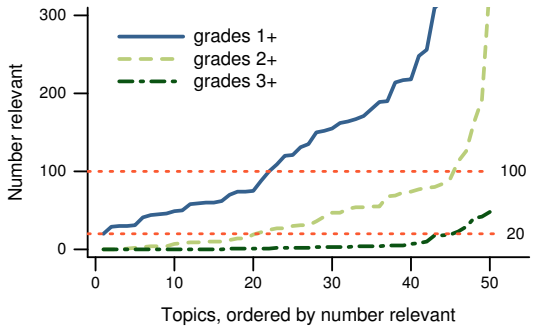

Fig. 1. Number of relevant documents in TREC 2013 Web Track judgments, using three different relevance thresholds. Relevance counts (on the vertical axis) are plotted in sorted order for each of the lines independently.

To create Figure 1, the number of judgments for each topic in the categories "1+", "2+", and "3+" were counted; each set of fifty numbers was sorted independently; and then the three sets of sorted values were plotted. As can be seen, only a small minority of the topics have more than twenty "Key" or better documents associated with them, and only thirty topics of the fifty topics have twenty or more "Highly relevant" (or better) documents associated with them. When computing AP@k and SSP@k the usual convention is to binarize the graded judgments, with "1+" regarded as "relevant" [13]. Figure 1 shows that SSP@100 underestimates AP@100 for more than 20 of the 50 TREC 2013 Web Track topics, in some cases by a factor of five.

Graded NDCG is based on numeric gain values rather than binary relevance. Suppose that the relevance category for document d is given by $q = r(d)$, where (in the case of the TREC 2013 data) $q \in \{0,1,2,3,4\}$. To convert relevance categories to real-valued gain amounts, one option is to use the mapping:

$$g(d) = \frac{2^{r(d)} - 1}{2^{\max\{r(d)\}}}\,. \tag{12}$$

If $\max\{r(d)\} = 4$, Equation 12 yields the mapping $g(\text{"Not"}) = 0.000$, $g(\text{"Relevant"}) = 0.063$, $g(\text{"Highly relevant"}) = 0.188$, $g(\text{"Key"}) = 0.438$, and $g(\text{"Navigational"}) = 0.938$.

Figure 2 shows the distribution of "Ideal DCG@20" values across the TREC 2013 Web Track topics. The two horizontal lines show two reference points: the scaling value $S_{20} \times (15/16)$ if it is assumed that there are $k = 20$ "Navigational" solutions for each topic; and another more conservative scaling value, computed assuming that each topic has one "Navigational" answer, and a further 19 "Key" answers (the line at 3.58). As can be seen, both of these reference lines are higher than the "Ideal DCG@20" value needed when computing NDCG@20, with the ratio between them variable across topics. That is, Figure 2 further supports our contention that SDCG@k and NDCG@k are different metrics and should not be conflated, especially when "across sets of topics" averages are being computed. Similar patterns occur in the TREC 2014 Web Track judgments.

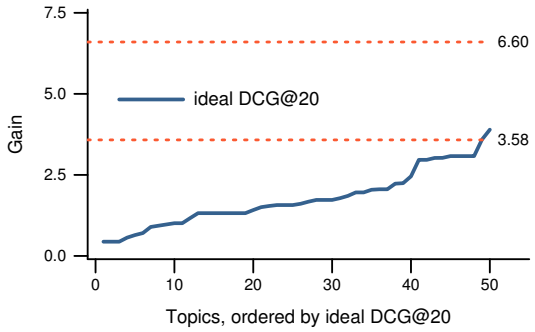

Fig. 2. "Ideal DCG@20" normalizing constants for NDCG@20 using the TREC 2013 Web Track relevance judgments. Per-topic gain values (on the vertical axis) are plotted in sorted order. The dotted lines represent two alternative scaling options for $k = 20$ evaluation.

5 Conclusion

We have examined the application of Tan and Clarke's MED technique to the recall-based metrics NDCG@k and AP@k. We observe that the mechanisms provided by Tan and Clarke and labeled by them as being MED-NDCG and MED-AP in fact relate to two different metrics, SDCG@k and SSP@k respectively. We also showed that MED(NDCG@$k,\cdot,\cdot$) and MED(AP@$k,\cdot,\cdot$) are both well-defined; and, moreover, that MED(NDCG@$k,\cdot,\cdot$) can be exactly computed in an efficient manner. We have also demonstrated by example and with reference to TREC relevance judgments that MED(NDCG@$k,\cdot,\cdot$), MED(AP@$k,\cdot,\cdot$), MED(SDCG@$k,\cdot,\cdot$) and MED(SSP@$k,\cdot,\cdot$) have different behavior and different outcomes; none of them is proportionately related in a predictable manner to the other three in any given situation, reinforcing the need for unambiguous nomenclature.

It is also possible to construct an AP-motivated version of the greedy process described in Algorithm 1, and use it to generate relevance assignments and hence effectiveness differences. The numerator weight for a free document y_i in the ith position of the ranking Y (required at step 6) is computed as:

$$wY[y_i] = \frac{1+|\{1 \le j < i \mid J[y_j] = 1\}|}{i} + \sum_{j \in \{i<j\le k | J[y_j]=1\}} \frac{1}{j}, \quad (13)$$

and is the increase in SP@$k(Y)$ that occurs if y_i is assigned a relevance of 1.0. The calculation of $wZ[z_i]$ is similarly changed, and then either R used as *denom* at steps 19 and 30, to match Equation 8; or $\min\{k,R\}$ used as *denom* to match Equation 9. Note, however, that the document weights and differences alter at each iteration, and need to be recomputed after each relevance label is assigned. In turn, that means that the sorting step is replaced by a sequence of up to k maximum-of-set computations over the differences in weights, taking $O(k^2)$ time. That is, Equation 13 describes the score adjustment for each as-yet unbound item, and the differences associated with all remaining items might alter at each step, making the computation more complex.

It is not claimed that the approach described by Equation 13 computes MED(AP@$k,\cdot,\cdot$), and it is presented purely as a heuristic that might provide relevance assignments that provide some guidance (in effect, an existence-based lower bound) as to what the MED(AP@$k,\cdot,\cdot$) value might be. We plan to explore this greedy AP-motivated MED mechanism in future work, with the alternative goals of either demonstrating that it correctly computes MED(AP@$k,\cdot,\cdot$), or, if counter-examples to that hypothesis are found, determining bounds on the extent to which values computed using Equation 13 can diverge from the true MED values. Nor is it clear how to generalize from the solution given here for NDCG to arbitrary recall-based metrics. For example, the Q-measure [14] is a weighted sum of an AP-like component and an NDCG-like component, and presents further challenges beyond even those associated with computing MED(AP@$k,\cdot,\cdot$). We make no proposal in regard to how MED(Q-measure@$k,\cdot,\cdot$) and similar recall-based metrics might be estimated or bounded, and leave this question as an open problem.

References

[1] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Trans. on Information Systems*, vol. 28, no. 4, pp. 20.1–20.38, 2010.

[2] L. Tan and C. L. A. Clarke, "A family of rank similarity measures based on maximized effectiveness difference," *IEEE Trans. on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 2865–2877, 2015.

[3] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Trans. on Information Systems*, vol. 27, no. 1, pp. 2.1–2.27, 2008.

[4] A. Moffat, P. Thomas, and F. Scholer, "Users versus models: What observation tells us about effectiveness metrics," in *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*, 2013, pp. 659–668.

[5] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*, 2009, pp. 621–630.

[6] A. Ashkan and C. L. A. Clarke, "On the informativeness of cascade and intent-aware effectiveness measures," in *Proc. Conf. on the World Wide Web (WWW)*, 2011, pp. 407–416.

[7] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[8] A. Moffat, "Seven numeric properties of effectiveness metrics," in *Proc. Asia Information Retrieval Societies Conf. (AIRS)*, 2013, pp. 1–12.

[9] S. Büttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[10] T. Sakai, "Alternatives to BPref," in *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, 2007, pp. 71–78.

[11] X. Lu, A. Moffat, and J. S. Culpepper, "The effect of pooling and evaluation depth on IR metrics," *Information Retrieval*, vol. 19, no. 4, pp. 416–445, 2016.

[12] W. Webber, A. Moffat, and J. Zobel, "Score standardization for inter-collection comparison of retrieval systems," in *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, 2008, pp. 51–58.

[13] K. Collins-Thompson, P. N. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees, "TREC 2013 web track overview," in *Proc. Text Retrieval Conf. (TREC)*, 2013.

[14] T. Sakai, "New performance metrics based on multigrade relevance: Their application to question answering," in *Proc. NII Testbeds and Community for Information Access Research (NTCIR)*, 2004.

Alistair Moffat completed a PhD at the University of Canterbury in 1986. Since then he has been a faculty member at The University of Melbourne, with interests in text and index compression, and algorithms for string search and information retrieval.

LIST OF TABLES

LIST OF FIGURES