Rank-Biased Quality Measurement for Sets and Rankings

Alistair Moffat
The University of Melbourne
Melbourne, Australia
ammoffat@unimelb.edu.au

Joel Mackenzie
The University of Queensland
Brisbane, Australia
joel.mackenzie@uq.edu.au

Antonio Mallia
Pinecone
New York, United States
antonio@pinecone.io

Matthias Petri
Amazon AGI
Los Angeles, United States
mkp@amazon.com

Abstract

Experiments often result in the need to compare an *observation* against a *reference*, where observation and reference are selections made from some specified domain. The goal is to determine how close the observation is to the ideal result represented by the reference, so that, all other things being equal, systems that achieve outputs closer to the ideal reference can be preferred for deployment. Both observation and reference might be sets of items, or might be ordered sequences (rankings) of items. There are thus four possible combinations between sets and rankings. Three of those possibilities are already familiar to IR researchers, and have received detailed exploration. Here we consider the fourth combination, that of comparing an observation set relative to a reference ranking. We introduce a new measurement that we call *rank-biased recall* to cover this scenario, and demonstrate its usefulness with a case study from multi-phase ranking. We also present a new top-weighted "ranking compared to ranking" measurement, and show that it represents a complementary assessment to the previous rank-biased overlap mechanism, and possesses distinctive characteristics.

CCS Concepts

• **Information systems** → **Retrieval effectiveness**.

Keywords

Evaluation; system comparison; precision; recall

ACM Reference Format:
Alistair Moffat, Joel Mackenzie, Antonio Mallia, and Matthias Petri. 2024. Rank-Biased Quality Measurement for Sets and Rankings. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24), December 9–12, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3673791.3698405

1 Introduction

Measurement is pervasive across all fields of science, often seeking to assess the quality of the results generated by some system when compared to what an ideal answer would be. That is, we compare an *observation* against a "perfect" (or "gold") *reference*, and compute a numeric score that measures the quality, or usefulness, of the system that generated the observation.

In many such measurement scenarios each of observation and reference are selections of items from some common domain. For example, an information retrieval (IR) system processes a query to generate a ranking of suggested documents. That answer list is next compared against knowledge of document relevance, and an effectiveness score computed; then we select (amongst alternative systems) the one with the highest score and say that it is the "best" system. In this measurement instance the observation is a *ranking*, an ordered sequence of documents; and the reference is a *set*, the known relevant documents.

A key observation that we make is that observation and reference can *each* be either an unordered set or an ordered ranking. This means that there are a total of four measurement scenarios that must be allowed for: set against set, set against ranking, ranking against set, and ranking against ranking. These four options are illustrated in Figure 1, with the notation "$X \mid Y$" meaning "an observation (of type) X is being assessed in the context of a reference (of type) Y".

Our project sits in this framework. We start by considering in detail the four possibilities that can arise when an observation B is to be compared to a reference R, noting that all four combinations arise in IR system evaluations. We then add:

- A new measurement technique addressing the "set | ranking" pairing, showing its relevance to IR via experiments on multi-stage retrieval systems (Section 3); and
- A new measurement technique addressing the "ranking | ranking" pairing, showing that it has properties that make it distinctive from previous top-weighted correlation coefficients (Section 4).

Bookending those two elements, Section 2 introduces the measurement taxonomy that we propose; and Section 5 concludes the presentation and offers suggestions for possible future work.

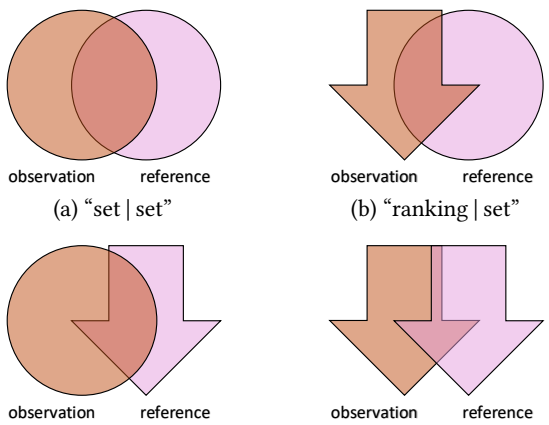
2 Background

This section reviews the ways that observations B (sets or ordered lists) have been compared to references R (sets or ordered lists).

"Set | Set" Measurement. Figure 1(a) is applicable if the observation (denoted B throughout) is a set; and a score is to be computed indicating how closely it approximates a reference that is also a set (denoted R). Boolean retrieval from an IR system, measured relative

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0724-7/24/12
https://doi.org/10.1145/3673791.3698405



(a) "set | set" (b) "ranking | set" (c) "set | ranking" (d) "ranking | ranking"

Figure 1: Four measurements scenarios that arise if the observations derived from some system are to be measured relative to a reference output. Circles are sets and arrows are rankings; and $X \mid Y$ means "observation X is to be evaluated relative to reference Y".

to a set of known relevant documents, is a classic instance of this "set | set" paradigm, with all elements in each of B and R having equal precedence. A standard response is to calculate precision via:

$$\text{Prec}(B \mid R) = \frac{1}{|B|} \sum_{i=1}^{|B|} (B_i \in R)\,, \tag{1}$$

where B_i is the ith item from B, and the expression "$B_i \in R$" is evaluated in a C-integer framework, and yields 0 if $B_i \notin R$, and yields 1 if $B_i \in R$. Precision measures "what fraction of B is of interest"; complementing it is:

$$\text{Recall}(B \mid R) = \frac{1}{|R|} \sum_{i=1}^{|R|} (R_i \in B)\,, \tag{2}$$

which is a measurement of "what fraction of R has been found". In particular, $\text{Prec}(B \mid R)$ and $\text{Recall}(B \mid R)$ both provide a score for the observation B, given the context established by the reference R. Our choice of notation allows an interesting symmetry to emerge:

$$\text{Recall}(B \mid R) = \text{Prec}(R \mid B)\,. \tag{3}$$

We will return to this duality in Section 3. Note that there is no requirement that $|B| = |R|$, and the circles in the Venn diagram in Figure 1(a) could have been drawn of different sizes.

"Ranking | Set" Measurement. Figure 1(b) illustrates another common IR evaluation situation, that of "ranking | set". Now the observations form an ordered ranking, with B_1 prioritized over B_2, and so on. Rankings arise from search services that incorporate a numeric similarity computation, and are presented as "top-k" output orderings, but understanding that each depth-k observation B is a prefix of a ranking that eventually includes every document in the collection.

In Figure 1(b) the reference R is still an unordered set of binary (or binarized) relevance judgments. With B a ranking rather than a set, the use of Prec() or Recall() is still possible if some fixed cutoff point k is used to form a prefix set from B. But "set | set" approaches do not reward systems that successfully place the relevant items near the top of the ranking, and a wide range of *top-weighted* measurements have evolved over the years. These include reciprocal rank (RR), which considers a prefix of B down to the first element that is also in R; average precision (AP) [3]; and normalized cumulative discounted gain (NDCG) [9].

In this work we focus on another "ranking | set" measurement, *rank-biased precision* [21], defined (for our purposes here) as:

$$\text{RBP}(B \mid R) = \frac{1-\phi}{\phi} \sum_{i=1}^{|B|} \left(\phi^i \cdot (B_i \in R)\right)\,, \tag{4}$$

with $|B|$ arbitrarily large, and with B_i always deemed to be "more visible" to the user than is B_{i+1}. Rank-biased precision reflects the aggregate experience of a population of users who consume the search results by always inspecting the first document, and then proceeding from one to the next with conditional continuation probability ϕ [21]. The convergence of the geometric sequence when $\phi < 1$ means that RBP can be computed even if $|B|$ is infinite, and that scores over monotonically longer prefixes converge to a limiting value [21]. Moffat et al. [22] consider a range of related mechanisms that are based on alternative user browsing models.

"Ranking | Ranking" Measurement. Figure 1(d) shows a third measurement scenario, when both B and R are rankings. One well-known approach is to use Kendall's τ, which calculates a score between -1 and $+1$ based on the fraction of times that B and R agree on the relative ordering of pairs of objects.

Webber et al. [30] consider "ranking | ranking" assessment in detail. They note that Kendall's τ has two potential drawbacks: first, it requires that B and R be permutations of each other (and hence also be of the same finite length); and second, it is not top-weighted. In particular, if B and R differ by a single swapped pair of adjacent elements, τ assigns the same score regardless of whether the swap occurs at the head of the two lists or at the tail.

Webber et al. go on to suggest an alternative computation they name *rank-biased overlap* (RBO) which addresses both of these concerns. Like RBP, RBO is controlled by a persistence parameter ϕ. What gets computed is an expectation of the overlap fraction observed by users who step down the two lists in tandem, continuing from depth i to depth $i+1$ with conditional probability ϕ. Define $B_{1..i}$ and $R_{1..i}$ to be the first i elements of B and R respectively, and with $B_{1..i} = B_{1..|B|}$ when $i > |B|$ and $R_{1..i} = R_{1..|R|}$ when $i > |R|$. Then:

$$\text{RBO}(B \mid R) = \frac{1-\phi}{\phi} \sum_{i=1}^{\infty} \frac{\phi^i}{i} \cdot |B_{1..i} \cap R_{1..i}|\,. \tag{5}$$

Note that $\text{RBO}(B \mid R) = \text{RBO}(R \mid B)$, and also that there is no requirement that B and R be permutations of each other or even of the same length. Webber et al. go on to consider a number of other somewhat technical issues, including noting that if prefixes of length i have an overlap of size v, then prefixes of length $i+1$ must have overlaps in the range v to $v+2$. These considerations allow lower and upper bounds to be computed for the infinite sum beyond depth $\min(|B|, |R|)$, thereby constraining the range of RBO scores that can emerge when B and R are finite prefixes of the infinite rankings assumed by Equation 5.

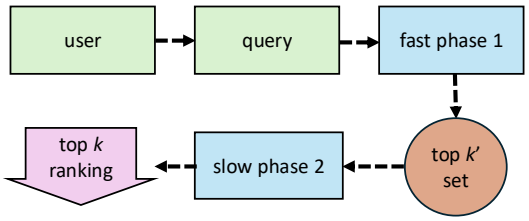


Figure 2: Assessing the quality of the first phase ranker in a multi-phase retrieval system. The top-k' list delivered to the second phase is a set B, whereas the final output is a top-k ranking R.

Other authors have also considered top-weighted correlation scores, notably Yilmaz et al. [31], who introduce a mechanism called τ_{AP} that is derived from Kendall's τ by considering disordered pairs, but biasing the pairs selected according to their positions in B. Webber et al. [30] provide further analysis of the τ_{AP} mechanism; one notable issue is that scores are not convergent as the rankings are extended, and nor is it symmetric. Other related work comes from Tan and Clarke [27] (see also Moffat [20]), who consider "ranking | ranking" measurement in terms of the maximum effectiveness difference (MED) that could be observed if both B and R were to be measured relative to a third set G of "gold" labels, where G might contain relevance judgments or might be empty.

3 Rank-Biased Recall

Now consider the "set | ranking" measurement scenario of Figure 1(c), an option that has had almost no attention in the literature.

An IR Application. Figure 2 explains why IR practitioners need to be interested in "set | ranking" measurement. Suppose that a retrieval system has multiple phases. The first is an efficient mechanism that determines a set of $k' > k$ elements that are passed to the less-efficient second (final) phase ranker [29]. That second phase ranker is an accurate but high-resource system that completely re-scores those k' documents, so as to identify and then present a final top-k subset. The question then is: how should first phase rankers be compared, when their role is to deliver a *set* of documents through to the second phase?

A similar area in which "set | ranking" evaluation is needed is shard-based or federated selective IR, where a decision is made as to which subset of documents shards is to be processed, and then the union of the retrieved documents from those selected shards becomes the set B for measurement purposes [10, 11, 25].[1]

Early Stage Assessment Is Not "Ranking | Ranking". If the first phase employs a numeric similarity computation, then the observation B is a prefix of a ranking. It might thus be tempting to apply a "ranking | ranking" mechanism, comparing that $|B| = k'$ ranking to the top-k ranking created by the second phase; or to a k'-item ranking generated by re-scoring all k' documents; or to a collection-wide ranking in which every document is scored using the second phase. For example, Mackenzie et al. [16] and MacAvaney and Tonellotto [14] measure quality using RBO; and Clarke et al. [4] and Mackenzie et al. [15] make use of Tan and Clarke [27]'s MED.

But such approaches give different scores to each permutation of B, creating a fake distinction between observations that become identical once re-scoring is applied.

Early Stage Assessment Is Not "Set | Set". Another approach that has been used to measure the quality of first phase rankers is to form a "true" top-k by applying the second-phase to the entire collection, and then taking the first k items from that global ranking be a set R. It is then natural to compute Recall$(B \mid R)$, the fraction of R that is present in the observation set B. This measurement approach is in use in a wide range of current experimental pipelines.

Different first stage rankers are compared by employing each to generate a top-k' set B for each query. A recall score is then computed for each query and each first stage system; and averaged over a large number of queries (possible because no relevance judgments are required) to obtain per-system average recall scores. Those scores are then compared to determine the best system.

But use of "set | set" measurement for early stage assessment has a clear drawback: not having the first element in R also present in B is likely to be more detrimental to perceived system quality than is omitting the k th. Yet recall treats both omissions equally. Similarly, replacing a missed document (in B) by the $k+1$ th element in R should be less disruptive to the score than including as a substitute the 99 th or 999 th document. That is, positions in R *do* have a bearing on the score, and do so even *beyond* rank k; whereas positions in B do *not* affect the score.

"Set | Ranking" Measurement. To address this clear gap we propose an approach that we call *rank-biased recall* (RBR):

$$\text{RBR}(B \mid R) = \frac{1-\phi}{\phi} \sum_{i=1}^{|R|} \left(\phi^i \cdot (R_i \in B) \right) . \tag{6}$$

Once the parameter ϕ has been chosen, each item in the reference ranking R is assigned a geometric weight, with R_1 having the highest weight, being $1-\phi$. The RBR score is then the sum of the weights of the elements that appear in B, a dual of the previous RBP computation. Table 1 gives a simple example, using a set B of size $k' = 5$, a reference sequence R of length 10, and $\phi = 0.6$. Having R_1 (document D07) present in B provides more than half of the final RBR score, which is calculated as (working from left to right in R, to reflect the ordering implied by Equation 6) $0.400 + 0.240 + 0.052 + 0.019 = 0.711$. Note how $R_7 =$ D06 also contributes to the RBR score, even though it is outside the top $k = 5$ in R. There is partial credit given for including in B something that was "nearly" in R's top-5, exactly as required.

The example in Table 1 includes an item $B_2 =$ D23 not present in R. From a pessimistic point of view it needs to be assumed that D23 (and in general, all items in $B \setminus R$) are located at the tail of any full ranking prefixed by R. That makes their ranks very large (shown as ∞ in the table), and their RBR contributions only marginally greater than zero. A more optimistic view is provided shortly.

As k' increases, additional items are included in B. As they are, the RBR score increases towards a limiting value of one, just as normal set-based recall also increases towards one as items are added to B. On the other hand, if $|B| = k'$ is regarded as being

[1] Noting that in some of these papers "RBR" stands for "Relevance Based Ranking". We will shortly overload the acronym, and ask the reader's understanding and forgiveness.

$R = \langle$D07, D04, D11, D12, D10, D15, D06, D22, D19, D28, $\ldots\rangle$					
i	1	2	3	4	5
B_i (e in Eqn. 8)	D06	D23	D10	D07	D04
$rank(R, e)$	7	∞	5	1	2
contrib., $\phi = 0.6$	0.019	*0.000*	0.052	0.400	0.240

Table 1: Computation of RBR. Reference R is an ordered list of arbitrary length (here, 10) of which the top-k are desired (here, 5), and the observation B is a set of size k' (here, 5); with (typically) $|R| \geq k' \geq k$. The RBR score of 0.711 is the sum of the contributions.

fixed, then RBR cannot be greater that $1 - \phi^{k'}$, which is achieved when $B = R_{1..k'}$, corresponding to both the perfect selection of the required set of size k, plus also perfect selection of near-misses out to position k' in R.

One pleasing consequence of the definition of RBR, and an elegant symmetry that mirrors the duality between precision and recall that was noted earlier (Equation 3), is that

$$\mathrm{RBR}(B \mid R) = \mathrm{RBP}(R \mid B)\,. \tag{7}$$

Note also that while Equation 6 indicates a summation over the elements in R, it is equally valid to sum over the elements in B, provided a mapping through to positions in R is available. In particular, if e is an element from the domain in question (for example a document number), and if $rank(R, e)$ is the rank position of e in R (and is ∞ if $e \notin R$), then we may equally write:

$$\mathrm{RBR}(B \mid R) = \frac{1-\phi}{\phi} \sum_{e \in B} \phi^{rank(R,e)}\,. \tag{8}$$

This version reflects the way that the example computation has been laid out in Table 1.

Residuals and Incomplete Information. In RBP there is a *residual* associated with the unseen tail of B, arising from two possible sources of inaccuracy [21]. The first is that the ranking B is a finite prefix of the full ranking, and if it is of length k, then a fraction ϕ^k of the probability distribution cannot be allocated. This is the *tail residual*. The second source of RBP inaccuracy comes from the incompleteness of R. Relevance judgments are rarely comprehensive, and B might include "unjudged" documents, as well as documents known to be in R (that is, have been judged relevant) and others known to not be in R (that is, judged non-relevant). In RBP the score ranges associated with both of these issues can be amalgamated together to provide a single residual in connection with the measured RBP value, providing an upper bound on how much the score might increase if either B was extended, or if R was made more comprehensive with regard to the universe of documents [21].

Rank-biased recall scores can also be subject to inaccuracy, but of a single variety. There is no ambiguity with regard to membership of the set B, since it is the observation. But R is typically represented by a finite prefix of a much longer sequence. That means that there may be elements in B for which $rank(R, B_i)$ is unknown, because B_i does not appear in the visible part of R. In the discussion above those ranks were taken to be ∞, so that the corresponding RBR terms were zero, a pessimistic accounting. On the other hand, the most *optimistic* thing that can happen is that each missing element appears in R just after the known prefix. If that happens, and there are $b = |B \setminus R|$ items in set B for which the rank in R is unknown, then the maximum possible increase in the RBR score is given by:

$$residual = \frac{1-\phi}{\phi} \sum_{i=1}^{b} \phi^{|R|+i}\,. \tag{9}$$

In the example in Table 1 we have $b = 1$ and $|R| = 10$, and hence have a residual of 0.002. That is, if R were to be extended beyond the 10 items shown in the table to eventually include D23, the final RBR score would lie between 0.711 and 0.711 + 0.002 = 0.713.

Choosing Parameters. Rank-biased precision is motivated by a user browsing model, and the choice of ϕ can be regarded as "selecting a type of user", either impatient (small ϕ) or patient (values of ϕ closer to one) [21]. Similarly, in RBO a value for ϕ is in part determined by the corresponding user model [30].

In the case of RBR it is less obvious how to select a suitable value for ϕ. The $\phi = 0.6$ discount employed in Table 1 is relatively aggressive, and suited primarily when k' and $|R|$ are small. For more typical values of k' in the range (say) 10 to 100, values of ϕ in the range 0.8 to perhaps 0.98 are more appropriate.

When selecting a value for ϕ, we suggest the following consideration: if $B1 = R_{1..k}$ is the best possible k-subset of R and obtains the highest score amongst all possible k-subsets, what fraction of that score should be awarded to $B2 = R_{k+1..2k}$? In other words, what relative score penalty should occur if the ideal k-subset is replaced by the next group of k consecutive elements from R?

Suppose that in response to that question we decide we would like $\mathrm{RBR}(B2, R) = f \cdot \mathrm{RBR}(B1, R)$ for some constant $0 < f < 1$. That implies:

$$f = \frac{\mathrm{RBR}(R_{k+1..2k} \mid R)}{\mathrm{RBR}(R_{1..k} \mid R)} = \frac{\phi(1-\phi)\sum_{i=k+1}^{2k} \phi^i}{\phi(1-\phi)\sum_{i=1}^{k} \phi^i} = \phi^k\,. \tag{10}$$

That is, ϕ^k can be interpreted as being the discount factor that applies if the best set of k items is replaced by the next-best non-overlapping set of k items. Conversely, if we have a value in mind for that discount factor, then $\phi = \sqrt[k]{f}$ should be chosen.

As an example, Table 2 supposes a reference list R of 10 items, and measures six different sets of observations (labeled $B1$ to $B6$, with memberships indicated by the "1" entries in the table), using two different values of ϕ. The first value of ϕ is chosen to give a $k = 3$ discount of $f = 0.5$, a relationship that can be verified by comparing the scores for $B1$ and $B4$. The second column of RBR scores uses a steeper discount, taking $f = 0.3$; again, compare the scores of $B1$ and $B4$. Note that it is possible for larger (but non-optimal) sets to yield higher RBR scores than smaller optimal sets – compare $B1$ and $B6$ for $\phi = 0.794$. But if ϕ is decreased to apply a steeper discount the RBR score relativities alter, and $B6$ no longer out-scores $B1$. Note also that the properties of the geometric distribution require that if the score for a k-shifted "next-best k-subset" is to be f times the score of the "best k-subset", then the "best k-subset" must get a score of $1 - f$, as shown by the two RBR scores associated with $B1$.

As further examples, if $k = 10$ and $f = 0.1$, then $\phi = 0.794$ should again be used; and $k = 100$ and $f = 0.05$ implies $\phi = 0.970$.

R	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	$\phi = 0.794$	$\phi = 0.669$
$B1$	1	1	1	-	-	-	-	-	-	-	0.500	0.700
$B2$	-	1	1	1	-	-	-	-	-	-	0.397	0.469
$B3$	-	-	1	1	1	-	-	-	-	-	0.315	0.314
$B4$	-	-	-	1	1	1	-	-	-	-	0.250	0.210
$B5$	-	1	-	1	1	1	-	-	-	-	0.414	0.431
$B6$	1	1	-	-	1	-	1	-	-	1	0.529	0.657

Table 2: A range of RBR scores, using either $\phi = \sqrt[3]{0.5} \approx 0.794$ or $\phi = \sqrt[3]{0.3} \approx 0.669$. In these examples all of the residuals are zero.

The flexibility to take into account elements from R outside the top-k and outside the top-k' is why we have chosen in Equations 6 and 8 to normalize by $(1-\phi)/\phi$ only, rather than also fold in a bound that might be imposed by $|B|$, which is finite. The definitions we have provided then allow observations B of different sizes to be directly compared, since the normalization is independent of $|B|$.

Finally in connection with Table 2, observe that if the six observations $B1$ to $B6$ were assessed via Recall@3, they would be assigned scores of 1.000, 0.666, 0.333, 0.000, 0.333, and 0.666 respectively; equating $B3$ and $B5$, and also equating $B2$ and $B6$. The RBR mechanism provides more nuanced scores that better differentiate between these pairs of observations, and at the same time provides the ability to directly control the extent of the top-weightedness, to cater for different measurement situations.

Ties. In "set | ranking" measurement the observation is a set, and all elements in it are regarded as being equally valuable. On the other hand, Equation 6 assigns decreasing weights to the elements in R, based on their ordinal position. But what if R is derived from a scoring process that permits ties? What if we have, for example, $R' = \langle\{\text{D07, D04, D11}\}, \{\text{D12}\}, \{\text{D10, D15}\}, \{\text{D06}\}, \{\text{D22, D19, D28}\}, \ldots\rangle$? How then should an RBR score be calculated in Table 1?

Our proposal is that tied elements be apportioned equal shares of their aggregate weight. For example, that would mean that each of D07, D04, and D11 have a weight in R' of $(0.400+0.240+0.144)/3 = 0.261$; and that at ranks 5 and 6 items D10 and D15 similarly share a weight of $(0.052+0.031)/2 = 0.041$. In turn, that would give the observation B in Table 1 a score of $0.019+0.000+0.041+0.261+0.261 = 0.583$, with the non-appearance of D11 and D15 in B being penalized more relative to R' than to R, because of their implicit promotion in the ranking of R' compared to the original R.

Experiment. To demonstrate the difference between use of RBR and Recall when evaluating first-phase retrieval systems, we carried out the following experiment. Starting with the 6,980 dev queries of the MSMARCO-v1 [1] passage collection, we applied a suite of traditional and learned sparse retrieval systems, following the setup of Mackenzie et al. [17]. The systems included a plain BM25 ranker [26]; a BM25 ranker with DocT5Query expansions [13, 23]; and four learned sparse systems including DeepImpact [2, 19], UniCOIL [8, 12] with DocT5Query expansions, UniCOIL with TILDE expansions [32], and SPLADEv2 [6, 7].

An index for each system was built with PISA [18] using the *anytime* extension [16] to allow for approximate retrieval. Each system was permitted to visit as many as $c = \{1, 2, 3, 4, 5, 10, 20, 50, 100\}$ document clusters, with $c = 100$ providing "safe" results, and lower values of c giving faster – albeit more approximate – results lists.

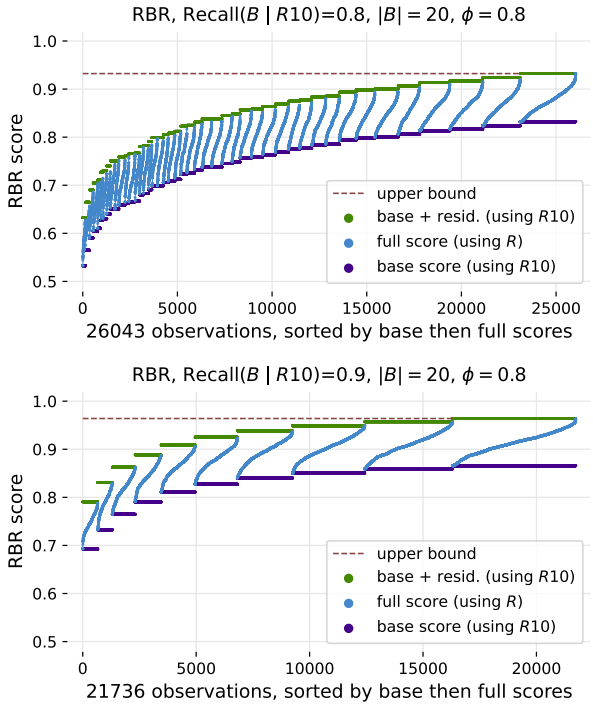


Figure 3: Rank-biased recall scores and ranges for a filtered set of observations B in which Recall$(B \mid R10) = 0.8$ (top) and Recall$(B \mid R10) = 0.9$ (bottom). Each observation B contains 20 documents, with $\phi = 0.8$ in both panes.

A reference ranking for each query was formed by pooling the sets of documents returned at depth 100 by the six $c = 100$ "safe" runs, and then re-scoring and re-ranking the pools using Mono-T5 [24] via the llm-rankers tool [33]. That is, we took Mono-T5 as being the second phase ranker, and for each query prepared a "full" reference ranking R containing between 100 and 600 documents. For the purposes of the experiment we then took $k = 10$ and formed a top-10 list for each query, $R10 = R_{1..10}$.

We next executed each of the 6×9 candidate first-phase systems on each query, generating observations of size $k' = |B| = 20$. Each of those sets B was then measured using Recall$(B \mid R10)$ to obtain a recall-based top-10 overlap score that could take on one of eleven different values 0.0, 0.1, to 1.0. As already noted, this is a common way of assessing the quality of a first-phase system. We then filtered the observations B, discarding all for which Recall$(B \mid R10) \neq 0.8$. That is, we formed a collection of observations of size $|B| = 20$ that all had exactly 8 elements in common with the corresponding "correct" top-10 list, and were thus judged by Recall@10 to all be of identical merit. That collection contained 26,043 observations. Finally, RBR with a parameter $\phi = 0.8$ was applied to each observation, using both the corresponding $R10$ reference ranking and also the extended ranking. That value of ϕ was selected in accordance with the commentary above.

The top pane in Figure 3 shows the RBR scores obtained, with the observations ordered by increasing RBR($B \mid R10$) base score. For example, the leftmost group of points is observations B in which the first two items in the ranking $R10$ were missed (there were 151); and the final group of points is observations in which the ninth and tenth items of $R10$ had been missed and replaced by others not in $R10$ (2,938 observations). In total, there are 45 groups, covering all possible combinations of two missed items. Residuals were also calculated relative to $R10$ (Equation 9). Because of the filtering these are constant, since there were always two items in B not appearing in $R10$ that at best could appear in positions R_{11} and R_{12}. The "base plus residual" lines in the plot show the maximum extent to which items not visible in the prefix $R10$ might affect the RBR score.

Within each of the 45 groups the observations were further sorted by their RBR($B \mid R$) "full" scores, noting that when $|R| \geq 100$ the residuals are effectively zero for $\phi = 0.8$. Finally, the top dashed line in the plot shows the best RBR score that can be attained by any observation of size $|B| = 20$ for which Recall($B \mid R10$) = 0.8.

By construction, all of the runs plotted in the top pane of Figure 3 have the same Recall@10 scores. However their RBR base scores vary considerably, reflecting the relative locations in $R10$ at which the two items have been missed. Our claim is that losing items one and two from R should be more detrimental than losing items nine and ten, and this is exactly the behavior visible across the "base score" line. Moreover, we further posit that even if it is the same two items that have been missed (say, items one and two), the assessment should be affected by the relative locations in R hosting the two replacement items. Substituting two missing items by the eleventh and twelfth ones from R should be less damaging to perceived quality than replacing them by items from positions (say) 234 and 345. The RBR($B \mid R$) "full score" lines in Figure 3 demonstrate exactly the required effect, with a wide range of RBR full scores evident, even within each group of equal RBR base scores. Those full scores generated from RBR($B \mid R$) must always lie in the range from base and base-plus-residual established by $R10$, but even within that range add important detail to the evaluation.

The bottom pane in Figure 3 shows the same experiment, but now with Recall($B \mid R10$) = 0.9, and thus ten groups of RBR base scores. Again RBR evaluated using the full ranking yields myriad fine-grained scores, whereas recall says "nothing to see here, these 21,736 observations are all the same". That is, both panes in Figure 3, derived from data generated by a plausible experimental context, show a level of gradated measurement that is simply not possible when using Recall. At risk of being simplistic, using recall for a "set | ranking" measurement is a bit like opening a nut with a sledgehammer.

Experimental Design. Given the developments we have described, how then should experiments on a two-stage system in which the first phase system is the "variable" be carried out? We suggest the following sequence of steps. First, suitable queries and documents should be obtained; and the desired second-stage ranker should be identified. As well, the length k of the required outputs that are to be generated by the overall combined system should be determined.

The second-stage ranker should then be executed in exhaustive "brute force" mode, to obtain for each query a reference ranking R of length substantially greater than k. For example, if $k = 100$, then the reference rankings should be perhaps 1000 long, so that the RBR residuals are small. Next, all possible first-stage rankers (comprising different systems and/or different parameter settings) need to be executed, to generate observations sets B of a range of sizes k'. Scores for RBR($B \mid R$) for each first-phase ranker, each k', and one or more suitable values of ϕ should next be calculated. As well, the execution time for each combination of k' and ranker should be captured.

With that data collected, the RBR scores can then be scatter-plotted against combined ranking time on the horizontal axis, one plot per choice of ϕ. In these plots the combined ranking time is the measured cost of the first-phase system for that system and k', plus the second-phase cost of re-ranking a set of k' items. In each such plot the points associated with each first-phase system describe a curve, parameterized by the sweep of k' values. The Pareto frontier of that set of lines then represents the subset of first-phase systems that are interesting for at least some combination of effectiveness and efficiency. First-phase systems that do not contribute to the Pareto frontier can be discarded.

Observation B	τ	RBO, ϕ =			RBA, ϕ =		
		0.6	0.7	0.8	0.6	0.7	0.8
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	1.00	1.00	0.99	0.97	0.99	0.97	0.89
[2, 1, 4, 3, 6, 5, 8, 7, 10, 9]	0.78	0.54	0.62	0.70	0.96	0.96	0.89
[5, 4, 3, 2, 1, 10, 9, 8, 7, 6]	0.11	0.23	0.33	0.46	0.78	0.86	0.85
[6, 7, 8, 9, 10, 1, 2, 3, 4, 5]	−0.11	0.04	0.10	0.22	0.51	0.68	0.77
[10, 9, 8, 7, 6, 5, 4, 3, 2, 1]	−1.00	0.04	0.10	0.22	0.40	0.60	0.73

Table 3: A "ranking | ranking" measurement for five permutations of $R = [1 \ldots 10]$.

4 Rank-Biased Alignment

In combination, RBP (Equation 4) and RBR (Equation 6) also suggest a new "ranking | ranking" measurement technique in the spirit of Figure 1(d). This section describes that mechanism, which we call *rank-biased alignment*, or RBA, and compares it to the previous RBO technique.

Rank-Biased Overlap. As was noted in Section 2, RBO calculates a weighted sum of overlap ratios across ranking depths.

Suppose that B and R are permutations, and have the same elements and are the same length. Table 3 lists five possible observations B, in the context of $R = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$, the identity permutation. The five permutations are scored by Kendall's τ, by RBO with three different values of ϕ, and by the rank-biased alignment measurement, which is described shortly. In the first row of the table, in which B and R are fully aligned, all of the measures provide high scores. At the bottom of the table B is the reverse of R, and the measures provide their smallest values. The three in-between sequences B have varying degrees of alignment with R, and are assigned correspondingly in-between scores.

However the second-to last row raises questions. In that row B is a piecewise rearrangement of R, with the top-5 items in order at the bottom, and the bottom-5 items swapped to the top (Figure 4(a)). Because Kendall's τ is not top-weighted, it assigns a score close to

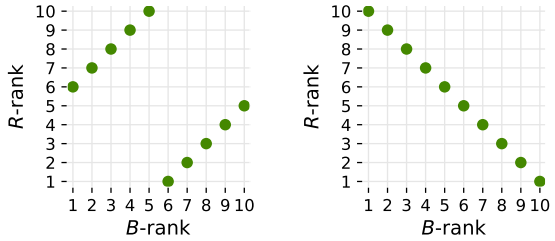


Figure 4: The last two rows of Table 3: (a) piecewise aligned; and (b) reversed. Fully aligned would run up the main diagonal.

zero. In contrast RBO assesses the same sequence as being highly disordered. Indeed, it gives it the same scores as for the reversed sequence in the last table row (Figure 4(b)).

It is not incorrect for RBO to assign the same scores (it occurs for all values of ϕ) to both of the permutations shown in Figure 4, since they have matching overlap counts at every depth. However, it could also be argued that the arrangement in Figure 4(a) should score more highly in a "ranking | ranking" sense than the one in Figure 4(b), because there is a higher degree of alignment between B and R.

Another issue that arises with RBO is computation of the residual – the extent to which the score might change if further elements of B and R become available to the measurement process. In particular, overlaps that first occur at depth d propagate to all depths greater than d. That means that an RBO score continues to increase even if all unseen tail elements in B are assumed to be disjoint from R. Webber et al. [30] provide formulae that allow the final RBO score range to be computed under different assumptions, but there is ambiguity in terms of what practitioners should do. In contrast, in RBP and RBR the computations at each depth are independent of any other depths, allowing simple closed formulations for score uncertainty; see, for example, Equation 9.

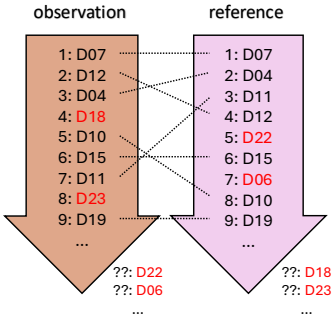
Combining RBP and RBR. In RBP a geometric weight is assigned to each item in B. In RBR the roles are switched, and a geometric weight is associated with each item in R. Given that context, our suggestion for "ranking | ranking" measurement is straightforward to motivate – we suggest that the geometric weight given to each item be influenced by its rank in both of B and R. Hence:

$$\mathrm{RBA}(B \mid R) = \frac{1-\phi}{\phi} \cdot \sum_{e \in B \cap R} \phi^{(rank(B,e)+rank(R,e))/2}. \qquad (11)$$

If B and R are fully aligned, the exponents of ϕ in the summation will be $[1, 2, 3, \ldots]$, and the RBA value will in the limit approach 1.0. On the other hand, elements that are misaligned between B and R give rise to larger-than-minimal exponents, and thus reduce the RBA score. Elements that occur in B but not in R, or vice-versa, do not contribute at all, and further reduce the score. In the worst case, if B and R are disjoint over the prefixes provided, then $\mathrm{RBA}(B \mid R) = 0$, matching the score that would be assigned by RBO.

Another corner case arises when B is the reverse of R:

$$\mathrm{RBA}(reverse(R), R) = \frac{1-\phi}{\phi} \cdot |R| \cdot \phi^{(|R|+1)/2}. \qquad (12)$$



Item	*avg_rank*	contrib.
D04	2.5	0.176
D06	∞	0.000
D07	1.0	0.300
D10	6.5	0.042
D11	4.0	0.103
D12	3.0	0.147
D15	6.0	0.050
D18	∞	0.000
D19	9.0	0.017
D22	∞	0.000
D23	∞	0.000

Figure 5: Example of RBA computation with $\phi = 0.7$. The RBA base score of 0.835 is the sum of the contributions.

This is the smallest permutation score, and corresponds to the bottom row in Table 3. As desired, RBA gives a higher score to the piecewise-swapped permutation considered in the second last row of Table 3, and measures that sequence as having some modest amount of alignment. Note also that

$$\mathrm{RBA}(B \mid R) = \mathrm{RBA}(R \mid B), \qquad (13)$$

and that the computation is symmetric.

Figure 5 gives an example RBA calculation for two non-conjoint rankings. The base score is the sum of the contributions listed, calculated from the average rank of the shared items, with four items not matched between the two rankings.

Residuals and Incomplete Information. Equation 11 gives a base RBA score. If no further common elements are encountered as B and R are extended, then no further increments will occur. That is, the score given by Equation 11 is a tight lower bound; again, a property that we regard as desirable.

It is also straightforward to compute a tight upper bound on an RBA score, given finite prefixes B and R (of possibly different lengths) of arbitrarily long rankings. In this case, elements of B that do not appear in the current prefix R will give rise to maximal contributions if they are assumed to occur in R in positions $|R| + 1$ through to $|B \cup R|$ in the order that they appear in B. Similarly, elements of R that do not appear in the current prefix B will give rise to maximal contributions if they occur in B in positions $|B| + 1$ through to $|B \cup R|$ in the order that they appear in R. Then, from position $|B \cup R| + 1$ onward, B and R should be assumed to be perfectly aligned. This is the tail residual associated with RBP, and contributes a further $\phi^{|B \cup R|}$ to the residual.

Algorithm 1 gives details, returning a *base* score that is the tight lower bound, and calculating the *extra* amount that arises if B and R are both extended to their union in a maximally-aligned manner. That upper RBA score bound is also sharp, and is attained if B and R are optimally completed out to an arbitrary length.

For example, in Figure 5 the residual based on the most advantageous positioning of the four red items is $(1-\phi)(\phi^{7.5} + \phi^{9} + \phi^{7} + \phi^{9.5}) = 0.097$. The tail residual contributes another $\phi^{11} = 0.020$.

Choosing Parameters. In the case in which the two rankings are permutations of each other, Equation 12 can be used to guide the

Algorithm 1 Computing RBA(B, R), for ordered observation list B and ordered reference list R. Lower and upper values are computed, assuming that B and R are finite prefixes of arbitrary rankings.

```
base, extra ← 0, 0
// first process the elements that appear in both B and R
for e ∈ B ∩ R, in any order do
    avg_rank ← (rank(B, e) + rank(R, e))/2
    base ← base + (1 − φ) · φ^(avg_rank−1)
// now process the elements that appear only in B
t ← |R| + 1
for e ∈ B \ R, in order from the head of B do
    avg_rank ← (rank(B, e) + t)/2
    extra ← extra + (1 − φ) · φ^(avg_rank−1)
    t ← t + 1
// and now process the elements that appear only in R
t ← |B| + 1
for e ∈ R \ B, in order from the head of R do
    avg_rank ← (t + rank(R, e))/2
    extra ← extra + (1 − φ) · φ^(avg_rank−1)
    t ← t + 1
return ⟨base, base + extra + φ^|B∪R|⟩
```

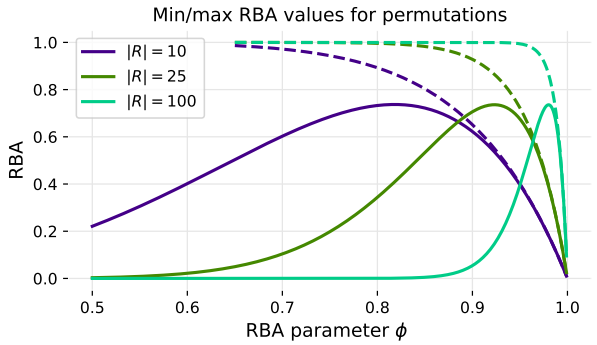


Figure 6: Range of RBA scores from minimum (reversed sequences, solid line of each pair) to maximum (fully aligned sequences, dashed lines) for permutations of three different lengths, as a function of ϕ.

choice of ϕ. For a given $k = |R| = |B|$, the practitioner first chooses – within the available range – the score that a reverse permutation should be assigned. That score then determines the ϕ that will attain it. Figure 6 plots minimum (Equation 12) and maximum (given by $1 - \phi^k$) RBA scores as a function of ϕ for three different values of k, and shows how the "reversed permutation" scores grow as ϕ increases. Figure 6 also shows that for each input length there is a value of ϕ above which the range of scores rapidly shrinks, and the use of RBA becomes unhelpful.

When B and R are not permutations and may instead be disjoint, the upper bounds shown in Figure 6 remain, but the lower bounds are zero. In this case, ϕ can be selected by considering a similar question as was posed in connection with RBR: if RBA$(B \mid R) = x$ what fraction $0 < f < 1$ of x score should be preserved if $2k$ completely different elements are introduced, k at the front of B and k at the front of R? Once this question is answered, the choice of $\phi = \sqrt[k]{f}$ is appropriate, as was discussed earlier.

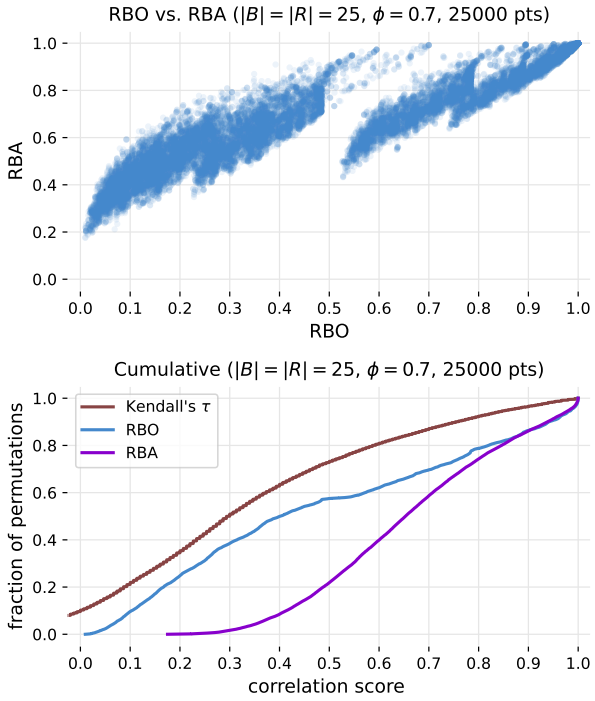


Figure 7: Distribution of RBA scores against RBO scores (top); and cumulative distributions for three correlations (bottom) for randomly generated high-correlation permutations. The relationships between the three sets of values are captured in Table 4.

Behavior on Random Permutations. We carried out the following experiment. Starting with identity permutations $B = R = [1..25]$, two randomly selected elements in B were swapped, representing a small perturbation relative to R. We then measured Kendall's τ, RBO using $\phi = 0.7$, and RBA also using $\phi = 0.7$. The same "perform a random swap" perturbation process was carried out a total of 24 more times, with B (in expectation) steadily diverging from R, and approaching random by the end of the sequence. That entire pattern, starting at $B = R$ and doing 25 random swaps one after the other, was carried out a total of 1000 times, to obtain 25,000 correlation scores, over a set of permutations that exhibits a bias towards being positively correlated.

Figure 7 shows the outcome, with the top pane plotting RBA scores (vertical axis) against RBO scores (horizontal axis) on a per-permutation basis (25,000 points plotted). The clumping of RBO values that is visible in that plot was a common outcome in this experiment, and is apparent, to some extent or another, across the spectrum of ϕ and k. Note the paucity of RBO scores in the vicinity of 0.5 – clearly, some parts of the RBO range are less likely to arise (at least, from this permutation generation regime). The lack of smoothness is also evident in the cumulative RBO distribution in the bottom pane of Figure 7. In contrast, the τ and RBA cumulative

Methods	$\phi = 0.7$	$\phi = 0.8$	$\phi = 0.9$
$\tau_b(\tau, \text{RBO})$	0.536	0.589	0.659
$\tau_b(\tau, \text{RBA})$	0.680	0.774	0.863
$\tau_b(\text{RBO}, \text{RBA})$	0.775	0.739	0.718

Table 4: Kendall's τ_b correlations for the data points generated for Figure 7 (left column) and for two other values of ϕ.

Obs. B	Ref. R	Example measurements
set	set	precision (Eqn. 1); recall (Eqn. 2)
ranking	set	rank-biased precision (Eqn. 4); plus many others
ranking	ranking	rank-biased overlap (Eqn. 5); rank-biased alignment (Eqn. 11)
set	ranking	rank-biased recall (Eqn. 6, Eqn. 8)

Table 5: Rank-biased approaches to measuring the quality of sets and rankings, to be interpreted in the context of Figure 1.

distributions are smoother, suggesting that the lack of smoothness is with RBO rather than the randomization process.

Table 4 takes the same data as was plotted in Figure 7 and computes pairwise correlation coefficients using Kendall's τ_b method, which takes into account the possibility of ordering ties. Each number in the table is a τ_b score computed from a total of 25,000 paired scores on a per-permutation basis. Unsurprisingly, the three approaches to measuring correlation – τ, RBO, and RBA – give scores that are themselves correlated (here on top-biased permutations, as described above, but also in general as well), with RBA seemingly located between τ and RBO in terms of the degree to which it is top-weighted. Also note that as ϕ is increased and the depth discount is weakened, both RBO (top row) and RBA (middle row) place increasing weight further down the rankings (which here are permutations) and hence shift closer to τ in their assessments.

From this experiment we can conclude that RBA has similar properties to RBO, but is (on a same-ϕ basis, at least) slightly less top-weighted. The spread of points in the scatter plot in Figure 7 also indicates that RBA is assessing top-weightedness differently to RBO, and hence that it is a complementary technique. There might also be some circumstances in which the smoother scoring distribution associated with RBA becomes an advantage.

5 Conclusion and Future Work

Table 5 summarizes the context and results of this paper, and links them back to Figure 1. We have added rank-biased recall to the toolkit of IR measurements, and in doing so have completed the fourth assessment combination between sets and rankings. Categorizing those four possible ways of measuring quality is also one of the key contributions of this work; and has led to a structure that has, in no small part, prompted the development of RBR. We hope that the reader shares our pleasure in the symmetry of what we have achieved in completing the roster of set against ranking quality assessment techniques.

As well, we have described some concerns with rank-biased overlap, and shown that the proposed alignment technique RBA – which itself came about as a blend of RBP and RBR – avoids them.

There remain areas where we do not yet have full understanding.

User Models. The previous rank-based precision and rank-biased overlap measurements for "ranking | set" and "ranking | ranking" assessments can both be described in terms of user models [21, 30], and hence can be argued as measuring the expected experience across a community of probabilistic users. The situation in regard to RBR and RBA is more complex, since neither of them connects the parameter ϕ with observable user behaviors. This challenge arises with many recall-based measurements, and occurs because in IR situations the user cannot be required to know how many relevant documents exist in the collection, and their perceptions of quality should ideally be based only on that which they have seen [21, 22], rather than that which they have not. For RBR we can imagine a population of users stepping though R, progressing from one item to the next with probability ϕ, and after each step, assessing how much of B has been encountered. But whether that is a plausible model is arguable. In particular, the "consumer" of the output from a first-phase ranker is another ranker, with no user involved. Maybe that provides an exemption from the need to describe a believable user model.

For RBA the situation is equally complex – our proposition here is that RBA measures an interesting quantity, but relating that conjecture through to behaviors that might be observable across a population of users remains an interesting challenge.

Nor have we addressed the various social, contextual, and commercial aspects of IR measurement considered by Thomas et al. [28] in their recent survey.

Future Work. There are many areas of our proposal that will benefit from further exploration. Corsi and Urbano [5] have developed detailed methodologies for handling ties in RBO, and their work might help do the same for RBA. More generally, it might be interesting to investigate other weightings for the items in R in conjunction with RBR, just as many alternative weightings have been developed within the "ranking | set" framework established by RBP. It might also be possible to carry our detailed experimentation – beyond the scope of what we have presented here – to more precisely characterize the differences between RBO and RBA, so that researchers and practitioners can be offered better guidance as to when each might be preferred.

Acknowledgment. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (project DP190101113) and a Google Research Scholar Grant. Lida Rashidi and Justin Zobel (University of Melbourne) participated in helpful early discussions. The referees provided useful feedback that has improved the paper.

Software. Implementations of the tools described in this paper are available at https://github.com/rankbiased/rbstar.

References

[1] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv:1611.09268v3*, 2018.

[2] S. Basnet, J. Gou, A. Mallia, and T. Suel. DeeperImpact: Optimizing sparse learned index structures. *arXiv:2405.17093*, 2024.

[3] C. Buckley and E. M. Voorhees. Retrieval system evaluation. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–78. MIT Press, 2005.

[4] C. L. A. Clarke, J. S. Culpepper, and A. Moffat. Assessing efficiency-effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments. *Inf. Retr.*, 19(4):351–377, 2016.

[5] M. Corsi and J. Urbano. The treatment of ties in rank-biased overlap. In *Proc. SIGIR*, pages 251–260, 2024.

[6] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv:2109.10086*, 2021.

[7] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant. Towards effective and efficient sparse neural information retrieval. *ACM Trans. Inf. Sys.*, 42(5): 116.1–116.46, 2024.

[8] L. Gao, Z. Dai, and J. Callan. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proc. NAACL-HLT*, pages 3030–3042, 2021.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.

[10] A. Kulkarni and J. Callan. Selective search: Efficient and effective search of large textual collections. *ACM Trans. Inf. Sys.*, 33(4):17.1–17.33, 2015.

[11] A. Kulkarni, A. S. Tigelaar, D. Hiemstra, and J. Callan. Shard ranking and cutoff estimation for topically partitioned collections. In *Proc. CIKM*, pages 555–564, 2012.

[12] J. Lin and X. Ma. A few brief notes on DeepImpact, COIL, and a conceptual framework for information retrieval techniques. *arXiv:2106.14807*, 2021.

[13] X. Ma, R. Pradeep, R. Nogueira, and J. Lin. Document expansions and learned sparse lexical representations for MSMARCO V1 and V2. In *Proc. SIGIR*, pages 3187–3197, 2022.

[14] S. MacAvaney and N. Tonellotto. A reproducibility study of PLAID. In *Proc. SIGIR*, pages 1411–1419, 2024.

[15] J. Mackenzie, J. S. Culpepper, R. Blanco, M. Crane, C. L. A. Clarke, and J. Lin. Query driven algorithm selection in early stage retrieval. In *Proc. WSDM*, pages 396–404, 2018.

[16] J. Mackenzie, M. Petri, and A. Moffat. Anytime ranking on document-ordered indexes. *ACM Trans. Inf. Sys.*, 40(1):13.1–13.32, 2022.

[17] J. Mackenzie, A. Trotman, and J. Lin. Efficient document-at-a-time and score-at-a-time query evaluation for learned sparse representations. *ACM Trans. Inf. Sys.*, 41(4):1–28, 2023.

[18] A. Mallia, M. Siedlaczek, J. Mackenzie, and T. Suel. PISA: Performant indexes and search for academia. In *Proc. OSIRRC at SIGIR 2019*, pages 50–56, 2019.

[19] A. Mallia, O. Khattab, N. Tonellotto, and T. Suel. Learning passage impacts for inverted indexes. In *Proc. SIGIR*, pages 1723–1727, 2021.

[20] A. Moffat. Computing maximized effectiveness distance for recall-based metrics. *IEEE Trans. Know. Data Eng.*, 30(1):198–203, 2018.

[21] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2, 2008.

[22] A. Moffat, J. Mackenzie, P. Thomas, and L. Azzopardi. A flexible framework for offline effectiveness metrics. In *Proc. SIGIR*, pages 578–587, 2022.

[23] R. Nogueira and J. Lin. From doc2query to docTTTTTquery, 2019. URL https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-latest.pdf. Unpublished report, David R. Cheriton School of Computer Science, University of Waterloo, Canada.

[24] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. In *Proc. EMNLP Findings*, pages 708–718, 2020.

[25] A. L. Powell and J. C. French. Comparing the performance of collection selection algorithms. *ACM Trans. Inf. Sys.*, 24(4):412–456, 2003.

[26] S. E. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trnd. Inf. Retr.*, 3:333–389, 2009.

[27] L. Tan and C. L. A. Clarke. A family of rank similarity measures based on maximized effectiveness difference. *IEEE Trans. Know. Data Eng.*, 27(11):2865–2877, 2015.

[28] P. Thomas, G. Kazai, N. Craswell, and S. Speilman. What matters in a measure? A perspective from large-scale search evaluation. In *Proc. SIGIR*, pages 282–292, 2024.

[29] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proc. SIGIR*, pages 105–114, 2011.

[30] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Sys.*, 28(4):20.1–20.38, 2010.

[31] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proc. SIGIR*, pages 587–594, 2008.

[32] S. Zhuang and G. Zuccon. TILDE: Term independent likelihood moDEl for passage re-ranking. In *Proc. SIGIR*, pages 1483–1492, 2021.

[33] S. Zhuang, H. Zhuang, B. Koopman, and G. Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proc. SIGIR*, pages 38–47, 2024.