Extending Rank-Biased Overlap (RBO) to Relevance Profiles

Thijs Houben

Supervisor: Julián Urbano[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

June 22, 2025

Name of the student: Thijs Houben
Final project course: CSE3000 Research Project
Thesis committee: Julián Urbano, Lilika Markatou

Extending Rank-Biased Overlap (RBO) to Relevance Profiles

Thijs Houben
Delft University of Technology
Delft, The Netherlands

Abstract

Rank-Biased Overlap (RBO) is a widely used metric for comparing ranked lists, due to its ability to handle incomplete and non-conjoint rankings while emphasizing top-ranked items. However, traditional RBO only considers the identity of ranked items, ignoring any associated relevance values. In many real-world applications, different systems may retrieve non-overlapping documents with similar informational value. This paper proposes an extension of RBO that incorporates graded relevance scores, enabling the comparison of rankings based on the information they convey rather than shared items alone.

Two relevance-aware variants for redefining RBO are proposed using cumulative gain. These variants are evaluated and analyzed using TREC ad hoc and simulated data, comparing them with each other and against standard RBO. The results demonstrate that the new RBO variants provide a more informative similarity measure when comparing rankings with differing identities but similar relevance patterns.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**; **Similarity measures**; *Retrieval effectiveness*; Information retrieval.

Keywords

Rank correlation, rank similarity, rank-biased overlap, relevance profile, graded relevance, gain

1 Introduction

Rankings are an integral part of modern everyday life: recommendations on streaming platforms, search engine results, feeds on social media platforms, and many more. The creation of such rankings is often heavily automated, using systems that aim to retrieve the most relevant items for the users need.

The field of information retrieval (IR) offers techniques to evaluate the rankings produced by these systems [11]. Traditional IR evaluation assigns a binary relevance score of 1 or 0 to each document retrieved based on some ground truth, indicating whether the items are relevant or not. Then, metrics such as recall, precision and F1 are calculated to assess how well the system retrieved relevant documents.

However, binary relevance is often too simplistic to reflect the differences in document usefulness. Graded relevance is an extension of binary relevance that captures the notion that some relevant items are more relevant than others. It accomplishes this by grading on a scale $R = [0, 1, ..., M]$, for some integer M[13].

While traditional binary relevance metrics have been adapted to handle graded relevance (e.g., generalized precision and recall, gP and gR [5]), several new metrics have been developed specifically for graded relevance. These include metrics like Rank-Biased Precision (RBP) and gain-based measures such as Cumulative Gain (CG), Discounted Cumulative Gain (DCG), and Normalized Discounted Cumulative Gain (nDCG) [4, 9].

A different class of metrics focuses on the similarity between rankings themselves. This is useful when comparing the outputs of different systems or tracking changes in a system over time. Classical rank correlation measures such as Kendall's τ and Spearman's ρ compare the orderings of the same set of items, but they assume full overlap between lists [6, 14].

However, this assumption is often violated in practical retrieval settings, which is why Webber et al. [16] introduced Ranked Biased Overlap (RBO). RBO uniquely combines three important properties. Firstly, RBO is able to compare rankings of lists that do not contain the same items, which is referred to as non-conjointness. Secondly, RBO is top-weighted, meaning it assigns higher similarity scores to lists that agree more on the highest-ranked items, even if they differ more on lower-ranked ones, compared to lists that align more at the bottom but diverge at the top. Finally, RBO is monotonic, which implies that comparing more items of two rankings, for instance the top 20 rather than the top 10, will never result in a lower score.

Classic RBO calculates its score based on the identities of the items in both rankings and would for instance return a score of 0 if the items between the rankings are disjoint. One could argue, however, from the IR perspective, that 2 rankings:

- Ranking S: ⟨doc1 (rel=3), doc2 (rel=2), doc3 (rel=0)⟩
- Ranking L: ⟨doc4 (rel=3), doc5 (rel=2), doc6 (rel=0)⟩

with disjoint items are equivalent, since the relevance of the documents at each position are the same.

Thus, the aim of this work is to adapt RBO, as described by Webber et al. [16], to provide a similarity score between rankings based on their profiles, represented by the relevance scores of items, rather than their identity. This is done by answering the question: *How can Rank-Biased Overlap (RBO) be extended for relevance values?*.

To answer this question and motivate the redefinitions, Section 2 first describes identity based RBO more in depth. This is then followed by Section 3, which proposes 2 relevance aware RBO redefinitions and Section 4, which evaluates the new metrics and compares them to identity based RBO. Finally, Section 5 summarizes the work and provides suggestions for further research, followed by Section 6, which discusses the integrity and reproducibility of the work.

2 RBO using identities

Throughout this paper, the notation in Table 1 will be introduced and used. Furthermore, Table 2 is an example of 3 rankings L, S and T, both in their identity and relevance score form. This example will be used to explain RBO and to motivate the choices made in later sections.

Table 1: Summary of notation

Symbol	Description
S, L	Rankings with prefixes of lengths s and l, where $s \leq l$
S_d^{rel}, S_d^{id}	Relevance and identity of items at rank d in S
$S_{n:m}$	(multi)-Set of items from rank n to m (inclusive) in S
d	Evaluation depth for computing agreement
X_d, D_d, A_d	Overlap, difference, and agreement at depth d
p	Persistence parameter of RBO
$CG_{S,d}$	Cumulative gain of ranking S at depth d
G_x	Gain of relevance value x
$\mathcal{N}$	Normalization factor
R	List of all relevance values
M	maximum relevance score in R
ϵ	Parameter for local agreement when one ranking has CG = 0

Table 2: Examples of rankings L, S and T in identity and relevance form, with ranking prefixes of length $l = 9$, $s = 5$ and $t = 4$. $R = [0, 1, 2, 3]$.

d	1	2	3	4	5	6	7	8	9	10	11	...
L^{id}	⟨e	p	q	c	f	a	b	h	y⟩	z	k	...
S^{id}	⟨a	p	e	z	i⟩	q	o	h	k	f	l	...
T^{id}	⟨r	o	e	z⟩	m	s	n	h	u	f	v	...
L^{rel}	⟨1	2	1	3	0	2	3	2	3⟩	3	0	...
S^{rel}	⟨2	2	1	3	0⟩	1	3	2	0	0	1	...
T^{rel}	⟨0	3	1	3⟩	0	2	3	2	1	0	3	...

Identity based RBO^{id}, as described by Webber et al. [16], is calculated using a weighted sum of similarity values between the two rankings at increasing depth, using the following formula:

$$\text{RBO}^{\text{id}}_{L^{\text{id}},S^{\text{id}},p} = \frac{1-p}{p}\sum_{d=1}^{\infty} A^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}\cdot p^d = \frac{1-p}{p}\Bigg(\underbrace{\sum_{d=1}^{s} A^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}\cdot p^d}_{1} + \underbrace{\sum_{d=s+1}^{l} A^{id}_{L^{\text{id}},S^{\text{id}},d}\cdot p^d}_{2} + \underbrace{\sum_{d=l+1}^{\infty} A^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}\cdot p^d}_{3}\Bigg). \tag{1}$$

The input consists of rankings L^{id} and S^{id} with prefix lengths l and s, and the persistence parameter p. L^{id} and S^{id} in Table 2 are examples of what the input rankings for original RBO might look like. The letters represent the identities of the documents, so, for instance, $L^{\text{id}}_1 = S^{\text{id}}_3 = e$ represents the same document in both rankings. Additionally, L^{id} has a prefix length of $l = 9$ and S^{id} of $s = 5$. This implies that during the RBO calculation, only the first 9 elements of L^{id} and 5 of S^{id} are known. It should also be noted that it is assumed that $s \leq l$.

Furthermore, The persistence p has a domain of $(0, 1)$ and is used to adjust the top-heaviness of RBO. A lower persistence has more weight for the similarity at the top of the rankings, while a higher persistence gives more weight to the similarity between lower ranks. Common choices for p are .8, .9 and .95, which models a user that compares the top 5, 10 and 20 documents respectively [3].

Using these inputs, Equation (1) calculates the RBO^{id} score by taking a weighted sum of the agreement $A^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}$, at increasing depth d. Agreement represents the ratio of the overlap, denoted as $X^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}$, between rankings L^{id} and S^{id} at depth d. It is calculated as

$$A^{\text{id}}_{L^{\text{id}},S^{\text{id}},d} = \frac{|X^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}|}{d} = \frac{|L^{id}_{:d} \cap S^{\text{id}}_{:d}|}{d}, \tag{2}$$

where $L^{\text{id}}_{:d}$ and $S^{\text{id}}_{:d}$ represent the set of the top d elements of L^{id} and S^{id}, respectively. For instance, again using the example in Table 2, $L^{\text{id}}_{:3} = \{e, p, q\}$, $S^{\text{id}}_{:3} = \{a, p, e\}$ and $A^{\text{id}}_{L^{\text{id}},S^{\text{id}},3} = \frac{2}{3}$. Also, for the sake of conciseness, henceforth the rankings are removed from the subscript, so $A^{\text{id}}_{L^{\text{id}},S^{\text{id}},d}$ is written as A^{id}_d, unless specific rankings are used as examples.

However, at depths $d > s$, calculating the overlap X^{id}_d as described in Equation (2) is problematic. This is because the rankings past the prefixes are unknown when calculating the RBO score and assumptions have to be made. Consequently, there is incomplete information about S^{id} when $s < d \leq l$ and about both S^{id} and L^{id} when $d > l$.

To highlight these different cases, Equation (1) is split up into 3 different parts. Part 1 handles the depth increments to depth s, which is the final depth at which there is a new known element in the prefix of S^{id}. Part 2 covers the depths between $s+1$ and l, where unseen elements are selected past the prefix of S^{id}, but there are still elements in the prefix of L^{id}. Finally, part 3 covers the depth past l, which represents the unseen parts of the complete rankings of the prefixes L^{id} and S^{id}. So, for the rankings L^{id} and S^{id} from the example in Table 2, part 1 goes until $d = s = 5$, part 2 until $d = l = 9$ and part 3 is for everything after.

Webber et al. [16] proposed three main ways to calculate part 2 and 3 of Equation (1). First of all, RBO_{MAX} calculates the maximum possible score by assuming an optimal continuation of the rankings past their prefixes. RBO_{MIN}, on the other hand, calculates the minimum possible score, by assuming the worst-case continuation past the prefixes. Finally, RBO_{EXT} uses the agreement between the prefixes of the rankings to give a point estimate for the final score.

3 Definitions for RBO using relevance

3.1 Definitions for base RBO

3.1.1 Motivation for cumulative gain. To redefine RBO to be relevance-based, ways to reformulate agreement are proposed, similar to the approach from Corsi and Urbano [3] when they extended RBO to be tie-aware. The following describes different possibilities for defining agreement function A^{rel}_d and why eventually a metric based on cumulative gain [4] was chosen.

Initially, one might think that RBO with relevance scores could be calculated simply through the use of the original agreement calculation as described in Equation (2). This does not work, however, since there are repeating elements. S^{rel} from Table 2, for instance, has the relevance score 2 at position 1 and 2, which for a normal set would result in $S^{\text{rel}}_{:2} = \{2\}$. As a result, there are cases such as

$S^{\text{rel}}_{:3} = L^{\text{rel}}_{:3} = \{1, 2\}$, which is undesirable, since the rankings are not the same up to $d = 3$.

Instead, $S^{\text{rel}}_{:d}$ could be defined as a multi-set. So, for the example, $S^{\text{rel}}_{:2} = \{2, 2\}$. Equation (2) as is could handle both sets and multi-sets.

Moreover, the distributions of the relevance scores in $S^{\text{rel}}_{:d}$ and $L^{\text{rel}}_{:d}$ could be compared using distribution based metrics such as KL divergence and Hellinger distance [2, 7].

What these approaches miss, however, is the overlap between different relevance scores. For instance, using the examples in Table 2, they would yield the same agreement for $A^{\text{rel}}_{L^{\text{rel}},S^{\text{rel}},1}$ and $A^{\text{rel}}_{T^{\text{rel}},S^{\text{rel}},1}$, which is 0. It is argued that the similarity between $L^{\text{rel}}_{:1}$ and $S^{\text{rel}}_{:1}$ is bigger than between $T^{\text{rel}}_{:1}$ and $S^{\text{rel}}_{:1}$, since a relevance score of 1 and 2 implies that the informational gain of the documents are more similar in the scope of a topic than a score of 0 and 2. Thus, this should also reflected in the agreement, which would not be the case with the aforementioned metrics.

Therefore, a different approach to define A^{rel}_{d} had to be taken and inspiration was drawn from preexisting graded relevance evaluation metrics. From those, cumulative gain based metrics were found to be the most promising. They provide higher scores for more similar relevance grades and accumulate gain up to and including rank d, where the gain value depends on the relevance assigned to the document [4].

Both DCG and nDCG have their own built in weights based on depth, which is unnecessary, since RBO already provides this, which leaves base cumulative gain.

This is a good choice, as it is emulates the overlap calculation of RBO^{id}, which does not care when items are seen, but just that they have been seen before or at rank d. Moreover, the agreement definition using cumulative gain, denoted as A^{CG}_{d}, provides the opportunity to assign scores such that $A^{CG}_{1,L^{rel},S^{rel}} \neq A^{CG}_{1,T^{rel},S^{rel}}$.

3.1.2 Definitions using cumulative gain. Agreement is reformulated using cumulative gain as follows:

$$A^{\text{CG,x}}_{S^{\text{rel}},L^{\text{rel}},d} = \frac{X^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d}}{\mathcal{N}^{\text{x}}_{S^{\text{rel}},L^{\text{rel}},d}} = 1 - \frac{D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d}}{\mathcal{N}^{\text{x}}_{S^{\text{rel}},L^{\text{rel}},d}}, \tag{3}$$

where $\mathcal{N}^{\text{x}}_{d}$ is a normalization factor for which 2 variants are proposed in Section 3.1.3. Furthermore,

$$X^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d} = \mathcal{N}^{\text{x}}_{S^{\text{rel}},L^{\text{rel}},d} - D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d} \tag{4}$$

and

$$D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d} = |CG_{S^{\text{rel}},d} - CG_{L^{\text{rel}},d}| \tag{5}$$

describe the overlap and difference between the cumulative gains CG_d, respectively. Since $D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d} = D^{\text{CG}}_{L^{\text{rel}},S^{\text{rel}},d}$, the redefinition is symmetric, just like RBO^{id}.

CG_d, as described by Järvelin and Kekäläinen [4], is defined as:

$$CG_{S^{\text{rel}},d} = \begin{cases} G_{S^{\text{rel}}_d}, & \text{if } d = 1, \\ G_{S^{\text{rel}}_d} + CG_{S^{\text{rel}},d-1}, & \text{otherwise,} \end{cases} \tag{6}$$

where $G_{S^{\text{rel}}_d}$ is the gain associated with the relevance value of the document in ranking S^{rel} at rank d.

This gain could be understood as the amount the document helps the user and is described by a gain function that maps the relevance of the document to a numerical gain value. In general, the gain function can map relevance scores to any number, as long as irrelevant documents have no gain, $G_0 = 0$, and the function is monotonically non-decreasing, $\forall x, y\ x \geq y \rightarrow G_x \geq G_y$.

There are some standard gain functions commonly used, however. The linear and exponential mappings each offer different ways to convert a document's relevance score into a gain value and they each have a hyperparameter θ.

- **Linear gain** increases linearly with the relevance score and is scaled by a factor θ. This approach assumes that each increase in relevance provides a consistent, incremental benefit [4]:

$$G^{\text{lin}}_{S^{\text{rel}}_d,\theta} = S^{\text{rel}}_{d} \cdot \theta, \tag{7}$$

- **Exponential gain** places more emphasis on higher relevance scores by applying an exponential transformation to the score. This is useful when the importance of more relevant documents increases rapidly with each level [1].

$$G^{\text{exp}}_{S^{\text{rel}}_d,\theta} = \theta^{S^{\text{rel}}_{d}} - 1, \tag{8}$$

Examples of these gain functions can be found in Table 3.

When 2 rankings are compared, it is assumed that the same gain function is used for both rankings and that the gain function is known. This is necessary to perform the normalization, which is discussed next.

Table 3: Examples of linear and exponential functions, based on the relevance scores from Table 2

d	1	2	3	4	5	6	7	8	9	10	11	...
$G^{\text{lin}}_{L^{\text{rel}}_d,1}$	⟨1	2	1	3	0	2	3	2	3⟩	3	0	...
$CG^{\text{lin}}_{L^{\text{rel}},d,1}$	⟨1	3	4	7	7	9	12	14	17⟩	20	20	...
$G^{\text{lin}}_{S^{\text{rel}}_d,1}$	⟨2	2	1	3	0⟩	1	3	2	0	0	1	...
$CG^{\text{lin}}_{S^{\text{rel}},d,1}$	⟨2	4	5	8	8⟩	9	12	14	14	14	15	...
$G^{\text{exp}}_{L^{\text{rel}}_d,2}$	⟨1	3	1	7	0	3	7	3	7⟩	7	0	...
$CG^{\text{exp}}_{L^{\textbf{rel}},d,2}$	⟨1	4	5	12	12	15	22	25	32 ⟩	39	39	...
$G^{\text{exp}}_{S^{\text{rel}}_d,2}$	⟨3	3	1	7	0⟩	1	7	3	0	0	1	...
$CG^{\text{exp}}_{S^{\text{rel}},d,2}$	⟨3	6	7	14	14⟩	15	22	25	25	25	26	...

3.1.3 Global and local maximum cumulative gain normalization. The final part of the agreement redefinition in Equation (3) that still has to be covered is the normalization factor $\mathcal{N}^{\text{x}}_{d}$, which has to ensure that $\frac{D^{\text{CG}}_{d}}{\mathcal{N}^{\text{x}}_{d}} \in [0, 1]$, since it must hold that $A^{\text{CG,x}}_{d} \in [0, 1]$. There are two methods, local and global normalization, proposed to achieve this.

First of all, local normalization normalizes the difference D_d^{CG} by dividing by the maximum cumulative gain of both rankings at depth d, and is defined as

$$\mathcal{N}^{\text{loc}}_{S^{\text{rel}},L^{\text{rel}},d} = max\{CG_{S^{\text{rel}},d}, CG_{L^{\text{rel}},d}\}. \tag{9}$$

Since $CG_{S^{\text{rel}},d} \geq 0$ and $CG_{L^{\text{rel}},d} \geq 0$, it must always be the case that $D_d^{\text{CG}} = |CG_{S^{\text{rel}},d} - CG_{L^{\text{rel}},d}| \leq max\{CG_{S^{\text{rel}},d}, CG_{L^{\text{rel}},d}\} = \mathcal{N}^{\text{loc}}_{S^{\text{rel}},L^{\text{rel}},d}$. Therefore, it holds that $\frac{D_d^{\text{CG}}}{\mathcal{N}_d^{\text{loc}}} \in [0, 1]$.

Local normalization has two problems, however. First of all, if $CG_{S^{\text{rel}},d} = CG_{L^{\text{rel}},d} = 0$, then $\mathcal{N}_d^{\text{loc}} = 0$ and as a result, D_d^{CG} would be divided by 0. In this case, since $CG_{S^{\text{rel}},d} = CG_{L^{\text{rel}},d}$, the agreement is set to 1.

Secondly, if $CG_{S^{\text{rel}},d} = 0$ and $CG_{L^{\text{rel}},d} \neq 0$ (or vice versa), then $\frac{D_d^{\text{CG}}}{\mathcal{N}_d^{\text{loc}}} = \frac{\mathcal{N}_d^{\text{loc}}}{\mathcal{N}_d^{\text{loc}}} = 1$. As a result, any $CG_{L^{\text{rel}},d} > 0$ would lead to $A_d^{\text{CG}} = 0$, which is unwanted. A CG_d of 1 is clearly closer to 0 than a CG_d of 2, and this should be reflected in the agreement.

To handle this specific instance, the following is used in case of one CG_d being equal to 0:

$$A_d^{\text{CG,loc}} = 1 - \frac{\mathcal{N}_d^{\text{loc}} - \epsilon + \epsilon \cdot \frac{\mathcal{N}_d^{\text{loc}}}{d \cdot G_M}}{\mathcal{N}_d^{\text{loc}}} = \frac{\epsilon}{\mathcal{N}^{\text{loc}}_{S^{\text{rel}},L^{\text{rel}},d}} - \frac{\epsilon}{d \cdot G_M}. \tag{10}$$

In this redefinition, an ϵ is subtracted from the difference $D_d^{\text{CG}} = \mathcal{N}_d^{\text{loc}}$, to ensure that agreement is not always 0. This can be interpreted as setting the CG_d that is equal to zero to ϵ. This epsilon should be bounded such that $\frac{\mathcal{N}_d^{\text{loc}} - \epsilon}{\mathcal{N}_d^{\text{loc}}} < \frac{\mathcal{N}_d^{\text{loc}} - \min_x\{G_x | G_x > 0\}}{\mathcal{N}_d^{\text{loc}}}$, which forces the constraint that the agreement between 0 and a nonzero CG_d is never greater or equal than that between the same CG_d and the lowest possible nonzero CG score, which is equivalent to the minimum nonzero gain value.

Furthermore, it should still be possible for the agreement to be 0, which should be the case when the CG scores are as distant as possible. This maximum distance can be represented as $d \cdot G_M$, where G_M represents the maximum gain, because M is the maximum relevance. It corresponds to the worst case where one ranking is filled with only 0's and the other with the maximum relevance M. Therefore, the term $\epsilon \cdot \frac{\mathcal{N}_d^{\text{loc}}}{d \cdot G_M}$ is added to D_d as well, since it simplifies to ϵ if $\mathcal{N}_d^{\text{loc}} = d \cdot G_M$, thus canceling out the ϵ's in Equation (10), again resulting in $A_d^{\text{CG,loc}} = 0$.

Due to this additional factor, the domain for ϵ can also include the minimum gain value, resulting in $(0, \min_x\{G_x | G_x > 0\}]$.

When all edge cases for local normalization are combined, the agreement is calculated as:

$$A^{\text{CG,loc}}_{S^{\text{rel}},L^{\text{rel}},d} = \begin{cases} 1, & \text{if } CG_{L^{\text{rel}},d} = 0 \wedge CG_{S^{\text{rel}},d} = 0, \\ \frac{\epsilon}{\mathcal{N}^{\text{loc}}_{S^{\text{rel}},L^{\text{rel}},d}} - \frac{\epsilon}{d \cdot G_M}, & \text{if } CG_{L^{\text{rel}},d} = 0 \vee CG_{S^{\text{rel}},d} = 0, \\ 1 - \frac{D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},d}}{\mathcal{N}^{\text{loc}}_{S^{\text{rel}},L^{\text{rel}},d}}, & \text{otherwise.} \end{cases} \tag{11}$$

To avoid this division into cases, a second normalization factor, called global normalization, is proposed:

$$\mathcal{N}^{\text{glo}}_{S^{\text{rel}},L^{\text{rel}},d} = d \cdot G_M. \tag{12}$$

It applies the same principal as the $\epsilon \cdot \frac{\mathcal{N}_d^{\text{loc}}}{d \cdot G_M}$ term in Equation (10), but then for the normalization of D_d^{CG}.

All in all, the agreement function $A_d^{\text{CG,x}}$, defined in Equation (3) using either the global normalization $\mathcal{N}^{\text{glo}}_{S,L,d}$ or the local normalization version from Equation (11), can be substituted into the base RBO formulation in Equation (1) in place of the identity-based agreement A_d^{id}.

Using this reformulation, the computation of part 1 of the RBO expression in Equation (1), which covers the known prefixes of the two rankings, is fully defined. However, parts 2 and 3, which involve reasoning over the continuation of the rankings beyond the observed prefix lengths, require additional assumptions and methods. The three main procedures for handling the continuation - RBO_{MIN}, RBO_{MAX} and RBO_{EXT} - are discussed in detail in the following Section.

3.2 Definitions for RBO_{MIN}, RBO_{MIN}, and RBO_{EXT}

3.2.1 Definition for RBO_{MIN}. After evaluating the prefix of ranking S^{rel}, RBO_{MIN} describes a tight lower-bound on the final possible RBO score. The following describes procedures to arrive at this minimum score for part 2 and part 3 of Equation (1), for both the local and global normalization factors. For part 2, an algorithm is provided and for part 3 there are closed form solutions.

Initially, for part 2, one might think that the least optimal way to continue S^{rel} past its prefix is to just add the maximum or minimum gain to $CG_{S^{\text{rel}},s}$, depending on whether it was smaller or larger than $CG_{L^{\text{rel}},s}$.

The reason this does not work is illustrated in the examples from Table 3. At depth 5, $CG_{S^{\text{rel}},d,1}$ is higher than $CG_{L^{\text{rel}},d,1}$. However, the remainder of the prefix of L^{rel} contains a large number of high relevance values, resulting in a lot of gain. If the maximum relevance value was continuously picked for S^{rel} for the remainder of the prefix of L^{rel}, its cumulative gain would be $CG_{S^{\text{rel}},l} = 8 + 4 \cdot 3 = 20$, which leads to a difference of $D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},l} = |20 - 17| = 3$.

If 0's had been picked, on the other hand, the difference would be $D^{\text{CG}}_{S^{\text{rel}},L^{\text{rel}},l} = |17 - 8| = 9$ and the weighted sum of intermediate agreements and future agreements, for most reasonable values of p lead to a lower RBO score. Furthermore, it could also be the case during part 2 that it is optimal to alternate multiple times which gain value is the highest. Due to such edge cases, all reachable cumulative gain values are explored.

This can be accomplished using Algorithm 1 (see Appendix A). It employs a dynamic programming approach to keep track of all CG values the ranking S^{rel} can reach for depths $[s + 1, l]$ in a map and the minimum RBO value that led to that CG. Then, once the map for $d = l$ has been created, the equations for part 3 are used to calculate the minimum possible RBO score.

Algorithm 1 uses multiple inputs, which represent the following. CG_S and CG_L are the cumulative gain scores at depth s. R is the list

with all possible relevance scores. G is the gain function used and *locnorm* is a boolean that is true if $\mathcal{N}_d^{\text{loc}}$ should be used and false if $\mathcal{N}_d^{glo}$ is used. The other parameters represent the same values as in the rest of the paper.

Table 4: Sets of reachable cumulative gain values at increasing depths. It shows that an arithmetic series like [0,1,2] expands more slowly than [0,9,99]

	s	s+1	s+2	s+3
[0, 1, 2]	{0}	{0, 1, 2}	{0, 1, 2, 3, 4}	{0, 1, 2, 3, 4, 5, 6}
[0, 9, 99]	{0}	{0, 9, 99}	{0, 9, 18, 99, 108, 198}	{0, 9, 18, 27, 99, 108, 117, 198, 207, 297}

There are some things to note about Algorithm 1. The space and time complexity is heavily dependent on the gain function, since having more duplicate CG values during an iteration leads to fewer values in the map. This results in a lower space complexity, as illustrated in Table 4. In the case of a linear gain function, the time complexity is $O(|R|^2 l^2)$ and the space complexity is $O(|R|l^2)$.

This is due to the fact that only $|R|-1$ new CG values are added to the map at each iteration. Using the example in Table 4 at depth $s+1$, $CG_{s+1} = 2$ can reach 3 gain values, of which 2 are new, $CG_{s+2} = 3$ and $CG_{s+2} = 4$. $CG_{s+1} = 1$, on the other hand, can reach $CG_{s+2} = 1$, $CG_{s+2} = 2$ and $CG_{s+2} = 3$, which were either already in the map or just explored by the highest CG value. As a result, the increase can be interpreted as the maximum CG value reaching $|R| - 1$ new CG values during each iteration.

For an exponential gain function with a θ larger than l, on the other hand, the total number of CG values that can be reached at depth d is described by $\binom{d-s+|R|-1}{d-s}$, since the only overlap that occurs is if the same gain values are picked, but in a different order. This is equivalent to the combinatorics problem of calculating the number of combinations with replacement for unordered lists [12]. Assuming for the worst case that $s = 0$, this leads to a space complexity of $O(\frac{(l+|R|-1)!}{l!(|R|-1)!})$ and a time complexity of $O(\frac{(l+|R|-1)!|R|}{l!(|R|-1)!})$.

For part 3, the calculation starts at depth $l + 1$, where D_l^{CG} is the final difference value that depends on the actual prefixes S^{rel} and L^{rel}. To minimize the agreement in part 3, the difference D_d^{CG} should be maximized.

At each depth increment, D_d^{CG} can increase at most by the maximum gain G_M, as it assumes the worst case continuation of the rankings, where the ranking with the lower CG score is assumed to only contain 0's for $d > l$ and the other the maximum relevance score. Therefore, the sum that minimizes part 3 can be described for $A_d^{\text{CG,glo}}$ as

$$\sum_{d=l+1}^{\infty} A_d^{\text{CG,glo}} p^d = \sum_{d=l+1}^{\infty} (1 - \frac{D_l^{\text{CG}} + G_M \cdot (d-l)}{G_M \cdot d}) p^d. \tag{13}$$

Using the fact, as described by Webber et al. [16], that

$$\sum_{d=1}^{\infty} \frac{p^d}{d} = \ln \frac{1}{1-p}, \tag{14}$$

the closed-form solution for Equation (13) can be derived:

$$\begin{aligned} \sum_{d=l+1}^{\infty} (1 - \frac{D_l^{\text{CG}} + G_M \cdot (d-l)}{G_M \cdot d}) p^d = \\ \frac{G_M \cdot l - D_l^{\text{CG}}}{G_M} \sum_{d=l+1}^{\infty} \frac{p^d}{d} = \\ \frac{G_M \cdot l - D_l^{\text{CG}}}{G_M} (\ln \frac{1}{1-p} - \sum_{d=1}^{l} \frac{p^d}{d}). \end{aligned} \tag{15}$$

For $A_d^{\text{CG,loc}}$, there are 2 main cases, due to the split in Equation (11). If neither ranking has a cumulative gain score of 0 at depth l, then part 3 can be described as

$$\begin{aligned} \sum_{d=l+1}^{\infty} A_d^{\text{CG,loc}} p^d = \sum_{d=l+1}^{\infty} (1 - \frac{D_l^{\text{CG}} + G_M \cdot (d-l)}{\mathcal{N}_l^{loc} + G_M \cdot (d-l)}) p^d = \\ \sum_{d=0}^{\infty} (1 - \frac{D_l^{\text{CG}} + G_M \cdot (d+1)}{\mathcal{N}_l^{\text{loc}} + G_M \cdot (d+1)}) p^{d+l+1} = \\ \frac{\mathcal{N}_l^{\text{loc}} - D_l}{G_M} p^{l+1} \sum_{d=0}^{\infty} \frac{p^d}{d + ((G_M)^{-1} \cdot \mathcal{N}_l^{loc} + 1)}. \end{aligned} \tag{16}$$

This contains a sum that is a Lerch transcendent [8], which can be rewritten as an integral and have the form:

$$\Phi(z, k, \alpha) = \sum_{n=0}^{\infty} \frac{z^n}{(n+\alpha)^k} = \frac{1}{\Gamma(k)} \int_0^{\infty} \frac{t^{s-1} \cdot e^{-\alpha t}}{1 - ze^{-t}} dt. \tag{17}$$

Filling in the corresponding values for the sum in Equation (16) results in:

$$\begin{aligned} \Phi(p, 1, (G_M)^{-1} \cdot \mathcal{N}_{S,L,l}^{loc} + 1) = \sum_{n=0}^{\infty} \frac{p^n}{n + (G_M)^{-1} \cdot \mathcal{N}_{S,L,l}^{loc} + 1} \\ = \int_0^{\infty} \frac{e^{-((G_M)^{-1} \cdot \mathcal{N}_{S,L,l}^{loc} + 1)t}}{1 - pe^{-t}} dt. \end{aligned} \tag{18}$$

Equation (18) does not have a closed-form solution with elementary functions. Consequently, RBO_{MIN} for local normalization has to be estimated. This can be achieved by evaluating the sum up to a depth at which the remainder of the sum is negligibly small, since the inside of the sum is decreasing. Alternatively, the integral can be estimated. This can, for instance, be done using numerical quadrature methods, such as adaptive Gauss-Kronrod integration [10].

When one of the rankings, or both, have a CG of 0, one CG score is kept at 0 and the other is increased by the maximum gain, resulting in:

$$\sum_{d=l+1}^{\infty} A_d^{\mathrm{CG,loc}} p^d = \sum_{d=l+1}^{\infty} \left(\frac{\epsilon}{\mathcal{N}_l^{loc} + G_M \cdot (d-l)} - \frac{\epsilon}{G_M \cdot d}\right) p^d = \frac{\epsilon}{G_M} p^{l+1} \sum_{d=0}^{\infty} \frac{p^d}{d + ((G_M)^{-1} \cdot \mathcal{N}_l^{loc} + 1)} - \frac{\epsilon}{G_M} \sum_{d=l+1}^{\infty} \frac{p^d}{d} = \frac{\epsilon}{G_M} p^{l+1} \Phi(p, 1, (G_M)^{-1} \cdot \mathcal{N}_l^{\mathrm{loc}} + 1) - \frac{\epsilon}{G_M} (\ln \frac{1}{1-p} - \sum_{d=1}^{l} \frac{p^d}{d}), \tag{19}$$

where similar derivations are used as for Equations (15) and (16), again resulting in a Lerch transcendent.

3.2.2 Definition for $\mathrm{RBO}_{\mathrm{MAX}}$. $\mathrm{RBO}_{\mathrm{MAX}}$ describes the upper bound of the score that can be reached, similar to how $\mathrm{RBO}_{\mathrm{MIN}}$ describes the lower bound. The following proposes an exact method for calculating the strict upper bound when linear gain is used, and highlights the complexity for the definition for other gain functions.

The main proposed procedure is to make $CG_{S^{\mathrm{rel}},d}$ and $CG_{L^{\mathrm{rel}},d}$ equal as quickly as possible. This is only guaranteed to work for gain functions that map to an arithmetic series, since they ensure that continually picking the values that minimize the CG at depth d always converges the final difference to 0.

However, if the gain function does not follow an arithmetic progression, choosing values that equalize the cumulative gains early may actually yield a lower RBO score than selecting values that keep the gains very close without ever matching exactly. This is due to RBO's top-weighted nature, which prioritizes small differences in cumulative gain at early depths. For instance, for $R = [0, 3, 5]$ $CG_{S^{\mathrm{rel,d}}} = 5$ and $CG_{L^{\mathrm{rel}},d} = 6$, $D_d^{\mathrm{CG}} = 1$ first has to be increased to $D_{d+1}^{\mathrm{CG}} = 2$ before it can be $D_{d+2}^{\mathrm{CG}} = 0$.

Moreover, in the case of $\mathcal{N}^{\mathrm{loc}}$, if the maximum gain value G_M is large, it may be preferable to initially keep D_d^{CG} constant and then add G_M to both rankings to induce a sharp increase in $\mathcal{N}^{\mathrm{loc}}$, thereby maximizing agreement. This is again not an issue for arithmetic series, since lowering the difference always leads to higher agreement scores than increasing the normalization factor.

Thus, the following procedure for part 2 only applies to gain values that map to an arithmetic series. The gain values are picked such that $CG_{S^{\mathrm{rel}},d}$ gets as close as possible to $CG_{L^{\mathrm{rel}},d}$, using

$$CG_{S^{\mathrm{rel}},d} = \begin{cases} CG_{S^{\mathrm{rel}},d-1}, & \text{if } CG_{L^{\mathrm{rel}},d} - CG_{S^{\mathrm{rel}},d-1} < 0, \\ CG_{L^{\mathrm{rel}},d}, & \text{if } 0 \le CG_{L^{\mathrm{rel}},d} - CG_{S^{\mathrm{rel}},d-1} < G_M, \\ CG_{\mathrm{rel}S,d-1} + G_M, & \text{otherwise.} \end{cases} \tag{20}$$

For similar reasons as part 2, for part 3 it is again assumed that the gain values form an arithmetic series. An exhaustive search, similar to $\mathrm{RBO}_{\mathrm{MIN}}$, is not possible due to part 3 going to infinity and thus it is left as an open problem.

In the case of arithmetic series, the maximum gain is added to the lowest CG value, while the highest CG value stays the same for the $k = \lfloor \frac{D_l^{CG}}{G_M} \rfloor$ ranks past the prefix of l. Then, at depth $d = k + l + 1$, the gain value is added that leads to $D_{k+l+1}^{CG} = 0$. From then on, $\forall d \ge k + l + 1 (A_d^{\mathrm{CG,x}} = 1)$. This results in the following redefinition for part 3 for both normalization factors $\mathcal{N}_d^{\mathrm{x}}$,

$$\sum_{d=l+1}^{\infty} A_d^{\mathrm{CG,x}} p^d = \sum_{d=l+1}^{l+k} (1 - \frac{D_l^{\mathrm{CG}} - G_M \cdot (d-l)}{\mathcal{N}_d^{\mathrm{x}}}) p^d + \sum_{d=l+k+1}^{\infty} p^d = \sum_{d=l+1}^{l+k} (1 - \frac{D_l^{\mathrm{CG}} - G_M \cdot (d-l)}{\mathcal{N}_d^{x}}) p^d + \frac{p}{1-p} - \sum_{d=1}^{l+k} p^d. \tag{21}$$

No additional formulations are required to handle the different cases of agreement in Equation (11) for local normalization. This is because the minimum CG score always increases and therefore there is no case where the CG is 0, unless $CG_{L^{\mathrm{rel}}} = CG_{S^{\mathrm{rel}}} = 0$, but then $A_d^{\mathrm{CG,loc}} = 1$ and the sum for perfect agreement for $d = l + k + 1$ can immediately be applied for $k = 0$.

3.2.3 Definition for $\mathrm{RBO}_{\mathrm{EXT}}$. For $\mathrm{RBO}_{\mathrm{EXT}}$, the point is to extrapolate the complete RBO score based on the prefixes S^{rel} and L^{rel} and provide a point estimate. For part 2 of Equation (1), the continuation of S^{rel} has to be extrapolated. For part 3 the extrapolation is for not only the rankings but also their agreement.

The extrapolation for the continuation of S^{rel} for part 2 is based on the CG observed up to the end of its prefix at depth s, and is calculated as

$$CG_{S^{\mathrm{rel}},d} = CG_{S^{\mathrm{rel}},d-1} + \frac{CG_{S^{\mathrm{rel}},s}}{s}. \tag{22}$$

For both normalization factors, calculating part 2 is equivalent to the calculation of part 1. The only thing that is adapted is the way $CG_{S^{\mathrm{rel}},d}$ is calculated.

For part 3, similar to the definition of Webber et al. [16], the agreement between S and L is assumed to stay the same after depth l, which gives

$$\sum_{d=l+1}^{\infty} A_d^{\mathrm{CG,x}} p^d = A_l^{\mathrm{CG,x}} \cdot \frac{p^{l+1}}{1-p}. \tag{23}$$

.

4 Experimental Evaluation

To analyze the behaviour of the proposed definitions, tests were performed using TREC (Text REtrieval Conference) run data from 2010-2014 in the ad hoc track [15], and simulated rankings based on the approach from Corsi and Urbano [3]. To compare the metrics, the $\mathrm{RBO}_{\mathrm{EXT}}$ score is used, with rankings of lengths up to 100. Furthermore, $p = 0.9$ and $\epsilon = \min_x\{G_x | G_x > 0\}$ are used.

The TREC data consists of rankings of at most 1000 documents for different topics, generated by different systems. Each year contains 50 different topics, resulting in 250 different topics in total. For each topic, a subset of the documents is given a graded relevance score, where a scale of [0,1,2,3] was used in 2010 and 2011, and a scale of [0,1,2,3,4] for 2012, 2013 and 2014. If the document was not given a relevance score for a certain topic, then it was assumed the relevance score was 0. Any ties that were present within the rankings have been broken at random, as this prevents the RBO score from being inflated [3].

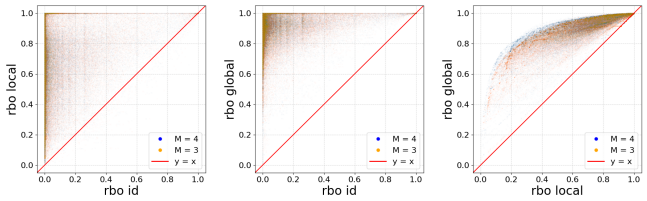


Figure 1: Comparisons between RBO^{id} and the two relevance based variants using different relevance domains, $\epsilon = 1$, G_1^{lin}, TREC data, p=0.9.

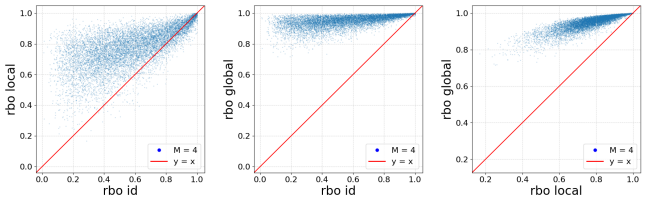


Figure 2: Comparisons between RBO^{id} and the two relevance based variants. $\epsilon = 1$, G_1^{lin}, Synthetic data, p=0.9.

To highlight the difference between the graded relevance scales of the TREC data before 2012 and after, the scatter plots use 2 different colors. The orange points in the scatter plots represent the comparison between rankings with documents graded using $R = [0, 1, 2, 3]$, while the blue points used $R = [0, 1, 2, 3, 4]$. In total, there are 110300 orange data points and 67350 blue data points for the TREC related scatter plots (Figure 1 and 4).

Moreover, 2 data sets of synthetic rankings were generated using an adaptation of the code from Corsi and Urbano [3] to create rankings that are more conjoined. It generates pairs of rankings of length 1000 based on a target Kendall τ. Both datasets are made up of 10000 pairs of rankings and grade the items using $R = [0, 1, 2, 3, 4]$. Due to differences in procedure used to generate the synthetic data, they are explained further in their respective Sections.

The Python and R code used to calculate RBO scores, generate synthetic rankings and produce graphs in this Section are available on Github[1]. Also, RBO^{id} was calculated using code from Corsi and Urbano [3].

4.1 Comparison between RBO scores using TREC data

Figure 1 compares the RBO scores for G_1^{lin} using TREC run data. The scatter plots reveal multiple properties of the different RBO metrics.

First of all, the scatter plot comparing global and local normalization shows that the variants are positively correlated. Also, the curve of blue points above the orange points reveals that the metrics are more similar when the maximum relevance is lower.

Furthermore, the scatter plots comparing to RBO^{id} show that there is no correlation between RBO^{id} and the relevance based metrics. For instance, when RBO^{id} finds little agreement, the relevance based metrics are still able to capture overlap. Thus, the relevance based metrics compute something different than RBO^{id}.

Moreover, in the vast majority of cases, the data points lie above the $x = y$ line, indicating that in general the relevance based metrics score higher than RBO^{id}. Sometimes, the local normalization is lower, which is further discussed in Section 4.3.

4.2 Comparison between RBO scores using synthetic data

Figure 2 uses synthetic ranking pairs that are more conjoint to compare the RBO scores. The pairs of rankings were generated using a target τ between 0.5 and 1 and the rankings were randomly truncated to a length between 10 and 100 [3]. Then, relevance scores were assigned to the items in each ranking pair, using the distribution $\{0 : 0.678, 1 : 0.212, 2 : 0.090, 3 : 0.018, 4 : 0.002\}$. This distribution was created by randomly sampling the first $10 \leq n < 100$ relevance scores of each 2014 TREC run.

Figure 2 shows similar trends as Figure 1, reaffirming that even when the identity-based rankings are more similar, the relevance based metrics still compute something different than RBO^{id}.

4.3 Analysis of the Effect of different relevance distributions using synthetic data

As stated at the end of Section 4.1, there are some cases where the RBO^{id} score is higher when compared to the local normalization RBO variant. This is also the case for the simulated data in Figure 2. These data points highlight a property of $A_d^{\text{CG,loc}}$, where $A_d^{\text{CG,loc}} = A_{d+1}^{\text{CG,loc}}$, when $S_{d+1}^{\text{rel}} = L_{d+1}^{\text{rel}} = 0$ and $CG_{d,S^{\text{rel}}} > 0$ and $CG_{d,L^{\text{rel}}} > 0$.

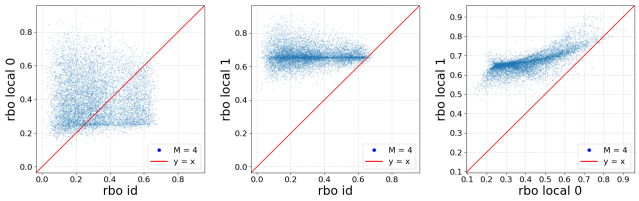
Figure 3 emphasis this using synthetic rankings [3]. Ranking pairs S^{id} and L^{id} are generated using a Kendall τ between 0.5 and 0.9 and truncated to a length of 100. The items in these rankings are then assigned a relevance score using the distribution $\{a : 0.92, b : 0.02, 2 : 0.02, 3 : 0.02, 4 : 0.02\}$, where a and b will be substituted out with 0 and 1. For instance, by substituting 0 for a and 1 for b, a is made the dominant relevance score. By swapping which element is dominant, two different pairs of S^{rel} and L^{rel} are created. For the final step, 2 new items f and g are introduced and assigned the relevance scores 1 and 4. f is then put at the top of S and g at the top of L. This sets $A_1^{\text{CG,x}} = \frac{1}{4}$ for both normalization factors, which ensures that the agreement not changing for local normalization is clear.

Figure 3 shows that the effect of altering the dominating element is different for both normalization variants. Figure 3 (a) depicts an increase of the RBO scores for local normalization, when the dominating relevance values is set from 0 to 1, which is as expected. Figure 3 (b), on the other hand, shows that for global normalization, the RBO scores that result from changing the dominating relevance are highly correlated, implying it is more robust to such changes.

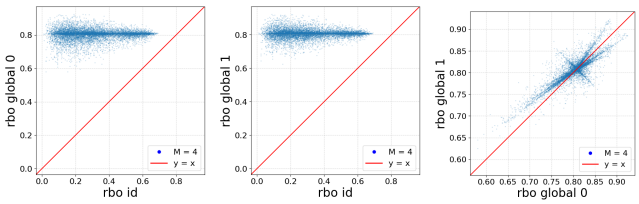
4.4 Effect of different gain functions using TREC data

The effect of different gain functions is presented in Figure 4 based on TREC data.

[1] https://github.com/ThijsH04/RBO_rel

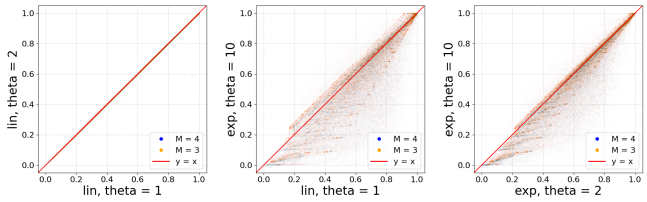


(a) Comparisons for local normalization

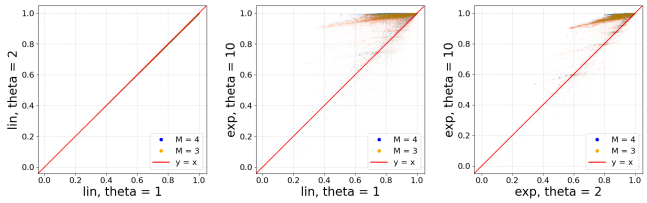


(b) Comparisons for global normalization

Figure 3: Comparisons between RBO$^{\text{id}}$ and relevance-based variants for synthetic data with $\epsilon = 1$,$p = 0.9$, $G^{\text{lin}}_{x,1}$, and $R = [0, 1, 2, 3, 4]$, $A_1 = \frac{1}{4}$. The first two columns show results for dominant relevance values 0 and 1, each assigned a probability of 0.92 (others received 0.02). The third column compares the relevance-based variants for dominant values 0 and 1.



(a) Comparisons for local normalization



(b) Comparisons for global normalization

Figure 4: Comparisons of RBO scores for local and global normalization using $G^{\text{lin}}_{x,1}$, $G^{\text{lin}}_{x,2}$, $G^{\text{exp}}_{x,2}$ and $G^{exp}_{x,10}$. TREC data, $\epsilon = 1$,p=0.9.

Figure 4 shows that for both normalization factors, different scalar terms θ for $G^{\text{lin}}_{x,\theta}$ lead to the same results. This is as expected, since the scalar is canceled out during the normalization.

Furthermore, Figure 4 also illustrates that exponential gain functions impact local and global normalization differently. In the case of global normalization (Figure 4(b)), using an exponential gain with a higher base θ results in higher RBO scores. This occurs because the global normalization factor scales with the maximum possible cumulative gain, which increases rapidly with higher θ. As a result, the relative differences between lower relevance scores become less significant, inflating the agreement and RBO score.

Conversely, in local normalization (Figure 4(a)), applying exponential gain with a larger θ leads to lower RBO scores. Since normalization in this case is based on the maximum cumulative gain observed between the two rankings at each depth, the denominator on increases a lot in the presence of high gain values. This makes any mismatch of high relevance more penalizing, by quickly reducing the agreement score.

Moreover, Figure 4 also shows that higher maximum relevance tend to inflate the RBO score of the global normalization variant, while local normalization is effected less.

5 Conclusions and Future Work

This work proposed a novel redefinition of Rank-Biased Overlap (RBO) [16] for relevance profiles, resulting in a similarity metric that accounts for graded relevance values associated with items in ranked lists. While traditional RBO compares rankings based solely on the identity of the items, the proposed reformulations capture similarity in informational value using relevance judgments of the items, even when two systems return disjoint sets of documents. This was achieved by redefining the agreement A_d to use the difference between the cumulative gain [4] of the rankings, while preserving RBO's key properties of top-weightedness, monotonicity, non-conjointness and symmetry.

Two redefinitions for A_d were presented and analyzed using real world (TREC) and synthetic data [3, 15]. The redefinitions use different ways to normalize the difference D_d between the cumulative gains of the rankings at depth d and have different use cases due to their strengths and weaknesses:

- **Global normalization**, normalizes the difference using the theoretical maximum difference at the depth of evaluation. Its scores tend to inflate when the maximum gain is higher, but it captures similarity between similar relevance's well.
- **Local normalization**, which normalizes the difference at each depth by the maximum cumulative gain observed in either ranking up to that point. It is less sensitive to higher maximum gain, but misses similarity with relevance's of 0.

The experiments also showed that the relevance-based RBO scores are generally higher than identity-based ones and that the scores are uncorrelated, highlighting that the redefinition calculate something different than original RBO.

Furthermore, using these definitions, an effort was made to reformulate the three RBO extensions that address uneven prefix lengths and the assumption of indefinite rankings. These formulations are complete for RBO_{MIN} and RBO_{EXT}.

For RBO_{MAX}, on the other hand, only complete procedures for linear gain functions were provided, leaving the redefinitions for exponential and in general arbitrary gain functions as an open problem.

Additional future work could also extend relevance-based RBO to handle ties, similar to recent work for identity-based RBO from Corsi and Urbano [3]. Furthermore, the proposed metrics could be compared with metrics such as nDCG and RBP [4, 9], or new relevance-based RBO reformulations.

6 Responsible Research

This section discusses the reproducibility and transparency of the work.

Throughout the paper, an effort was made to make the thought process behind decisions clear. The main goal of the paper was to explore possible redefinitions for RBO for ranking profiles and we were never under the assumption that the proposed solutions are perfect (the definitions are incomplete after all). By highlighting the thought process of the decisions made and the resulting flaws and short, future research on this topic, by for instance our responsible professor, can challenge our reasoning, possibly resulting in better redefinitions.

Furthermore, reproducibility of the work was kept in mind during the creation of the paper. By making the code, synthetic data, and results open source on Github[2], a reader should be able to reproduce and verify our results by following the descriptions in the Jupyter notebook. The aim was to make the Jupyter notebook easy to follow, to lower the burden of reproducing the results.

However, one possible issue with the reproducibility, is that the TREC dataset used is not present in the repository, due to its license. We assume this not to be a major problem, since the assumption is made that a researcher who aims to reproduce the results most likely is familiar with the TREC dataset and has access to it. Moreover, the Jupyter notebook makes it clear in which directory to store the TREC dataset to produce the required output.

As for the transparency of the work, ideas and code from literature were cited to the best of our ability. However, pointers given by the supervisor, or ideas that came up during discussions with the supervisor and peers, are not formally attributed, as there may only be one author for this thesis.

References

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 89–96. doi:10.1145/1102351.1102363

[2] L.L. Cam. 2012. *Asymptotic Methods in Statistical Decision Theory*. Springer New York. https://books.google.nl/books?id=L4juBwAAQBAJ

[3] Matteo Corsi and Julián Urbano. 2024. The Treatment of Ties in Rank-Biased Overlap. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 251–260. doi:10.1145/3626772.3657700

[4] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. doi:10.1145/582415.582418

[5] Jaana Kekäläinen and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *J. Am. Soc. Inf. Sci. Technol.* 53, 13 (Nov. 2002), 1120–1129. doi:10.1002/asi.10137

[6] Maurice Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* (1938).

[7] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86. http://www.jstor.org/stable/2236703

[8] M. Lerch. 1887. Note sur la fonction $\mathfrak{K}(w,x,s) = \sum_{k=0}^{\infty} \frac{e^{2k\pi ix}}{(w+k)^s}$. *Acta Math.* 11 (1887), 19–24. doi:10.1007/BF02612318

[9] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (Dec. 2008), 27 pages. doi:10.1145/1416950.1416952

[10] R. Piessens. 1983. *Quadpack: A Subroutine Package for Automatic Integration*. Springer-Verlag. https://books.google.nl/books?id=m2AZAQAAIAAJ

[11] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann.

[12] Kenneth H. Rosen. 2006. *Discrete Mathematics and Its Applications: And Its Applications*. McGraw-Hill Higher Education.

[13] Eero Sormunen. 2002. Liberal relevance criteria of TREC -: counting on negligible documents?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland) *(SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 324–330. doi:10.1145/564376.564433

[14] C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72–101. http://www.jstor.org/stable/1412159

[15] Ellen M. Voorhees and Donna K. Harman. 2005. TREC: Continuing Information Retrieval Evaluation. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*. National Institute of Standards and Technology (NIST).

[16] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (Nov. 2010), 38 pages. doi:10.1145/1852102.1852106

[2] https://github.com/ThijsH04/RBO_rel

A Algorithm for RBO_{MIN}

Algorithm 1 Algorithm for calculating RBO_{MIN} for part 2 & 3

procedure RBOMIN($CG_S, CG_L, R, s, l, L, p, G, locnorm$)
  $CGMap \leftarrow$ `empty map`
  $CGMap[CG_S] = 0$
  $d \leftarrow s + 1$
  while $d \leq l$ **do**
    $NextCGMap \leftarrow$ `empty map`
    $CG_L \leftarrow CG_L + G(L_d)$
    for $r \in R$ **do**
      for $cg \in CGMap$ **do**
        $NewCG \leftarrow cg + G(r)$
        $RBOScore \leftarrow CGMap[cg] + agreement(NewCG, CG_L, R, G, locnorm) \cdot p^d$
        if $NewCG \in NextCGMap$ **then**
          $NextCGMap[NewCG] \leftarrow max\{NextCGMap[NewCG], RBOScore\}$
        else
          $NextCGMap[NewCG] \leftarrow RBOScore$
        end if
      end for
    end for
    $CGMap \leftarrow NextCGMap$
    $d \leftarrow d + 1$
  end while
  $MinScore \leftarrow \infty$
  for $cg \in CGMap$ **do**
    $MinScore \leftarrow min\{MinScore, CGMap[cg] + part3(R, CG_L, cg, l, p, G, locnorm)\}$
  end for
  return $MinScore$
end procedure

Received 22 June 2025