On the Extrapolation of Rank-Biased Overlap and the Assumption of Constant Agreement

Konstantin-Asen Yordanov[1]
Supervisors: Julián Urbano[1]**, Matteo Corsi**[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 23, 2024

Name of the student: Konstantin-Asen Yordanov
Final project course: CSE3000 Research Project
Thesis committee: Julián Urbano, Matteo Corsi, Mathijs Molenaar

Abstract

As a point estimate of the similarity score between two possibly indefinite rankings, extrapolated rank-biased overlap (RBO_{EXT}) uses the assumption that the agreement observed at the last evaluation depth continues indefinitely across the unseen tails of the two lists. This assumption does not account for any patterns that occur in the visible prefixes, imposing a strict restriction on the extrapolation. In an effort to improve the accuracy of RBO_{EXT}, three reformulations with a relaxed theoretical basis are proposed in this paper: one continually re-uses the agreement from the previous depth while the other two rely on regression to fit a function on the seen agreements. Using synthetic data, the performance of these new extrapolation methods is compared to the original's in terms of closeness to the true RBO score as well as the average distance between assumed and actual agreement in the rankings' unseen tails. Overall, an impactful difference is observed in the estimates of agreement generated by the four approaches: as the trends from the visible prefixes are barely captured by the simpler techniques or closely-reproduced by the more flexible ones, the trade-off between under- and overfitting becomes increasingly relevant. The results thus indicate a need for some middle-ground to be established such that it factors in the observed patterns while also generalizing well for the tails.

1 Introduction

In many scenarios where results are aggregated from the execution of an operation – such as a search engine running a user-submitted query – rankings are produced. The latter are usually top-weighted, meaning that the ordering of elements at the top of the list matters more than that at lower positions, and incomplete as they do not contain all possible values and instead only cover a subset of the full domain [Webber *et al.*, 2010]. Additionally, rankings are often indefinite, which necessitates truncating them in a context-dependent and therefore inconsistent manner to produce finite prefixes whose exhaustive traversal would be feasible.

In practice, rankings are typically evaluated via relevance judgements which are provided by human assessors and thus inherently subjective, varying greatly depending on what criteria for usefulness are considered [Bar-Ilan, 2005]. Meanwhile, to compare and contrast the algorithms which generate rankings, computation-based similarity measures that do not rely on opinion-polling can be applied on the observed orderings, yielding insights about correlation and overlap, domain coverage, and information entropy [Cardoso and Magalhães, 2011]. However, due to the feasibility constraints imposed by indefinite rankings and the absence of a "gold-standard" ranking (or "ground truth") in most cases, only the items ranked at the top k positions by any two given algorithms are considered [Fagin *et al.*, 2003]. This makes it desirable for comparison measures to output a similarity estimate for the full lists even when only partial prefixes are available.

Satisfying precisely this requirement, rank-biased overlap (RBO) is a top-weighted, overlap-based similarity measure that bounds the full similarity score based on an evaluation of the two rankings' visible prefixes [Webber *et al.*, 2010]. The greater the length of the lists provided as input, the narrower the bounds become, and the more accurate the estimation. To capture the property of top-weightedness, RBO is parameterized by p (persistence), which represents a user's probability of proceeding further down the two lists and considering the items ranked next. Furthermore, as it is based solely on set intersection and a convergent sum of geometrically-decreasing weights that are fixed per rank, RBO provides consistent similarity scores even when the given rankings are non-conjoint, incomplete, or indefinite [Webber *et al.*, 2010].

1.1 Motivation and Aim of the Study

In most cases where RBO is applied, a point estimate for the full similarity score (RBO_{EXT}) is usually computed. To do so, the last-observed agreement at evaluation depth k is assumed to continue indefinitely [Webber *et al.*, 2010]. This, however, is quite a strong assumption as it does not account for specific patterns throughout the visible prefixes. Consequently, there is no way of introducing variability into the agreement-values for the unseen tails – which makes the original extrapolation approach very inflexible. To address precisely that issue, this study investigates the following research question:

> *How does redefining extrapolated RBO by altering the assumption of constant agreement for elements in the unseen sections of the two rankings influence the accuracy of the RBO point estimate?*

Therefore, by proposing alternative formulations of the single RBO score and then performing a comparative evaluation, the study aims to determine if a more accurate point estimate can be achieved and thereby confirm whether the modification of RBO_{EXT} merits further research.

1.2 Contributions

To redefine RBO_{EXT}, this study is centered on the interpretation of agreement as the probability that a randomly-selected element appears in both rankings [Webber *et al.*, 2010]. This probability can be estimated and assigned to each unseen item as the latter's degree of membership (or contribution). In turn, rather than extrapolating out from a single agreement-value, both the assumed agreement and the RBO point estimate can be computed iteratively at each depth, starting right after the end of the shorter visible prefix and continuing indefinitely.

Three approaches for carrying out the latter procedure are presented in this paper. The first newly-proposed formulation simply re-uses the assumed agreement at the previous depth as the estimated probability of membership for every unseen item. Meanwhile, the other two rely on regression techniques that fit a function on all observed agreements up to the end of the shorter visible prefix and output an estimate of the membership probability. To account for the trade-off between interpretability and complexity vis-à-vis the choice of a regression model, one approach involves logistic regression while the other uses the more flexible generalized additive logistic model (Logistic GAM) [Hastie and Tibshirani, 1986].

To compare the performance of these three reformulations against the original RBO$_{\text{EXT}}$ implementation, simulated rankings are generated and exhaustively traversed to compute the real RBO and agreement-values, which serve as the objective gold standard. Defined as closeness to the latter, accuracy is the performance measure based on which the four RBO point estimates and their corresponding agreement-approximations are evaluated. Thus, the differences between the four variants of RBO$_{\text{EXT}}$ are illustrated, and based on them, it is determined whether a redefinition of the single RBO score could offer an improvement over the original framework and should remain an object of future studies.

This paper has the following structure: Section 2 elaborates on background concepts related to rankings and outlines the relevant properties of RBO. Section 3 then demonstrates the mathematical framework behind the three proposed reformulations of RBO$_{\text{EXT}}$, followed by an explanation of the experimental setup and a presentation of the results in Section 4. Section 5 offers a discussion of the observed results, placing them in a broader context. Afterwards, Section 6 summarizes the findings, highlights the main conclusions, and offers recommendations for further research. Finally, Section 7 tackles the study's ethical aspects and the methods' reproducibility.

2 Background

This section establishes the theoretical framework for the rest of the paper, presenting the relevant background concepts regarding rankings and the RBO similarity measure. First, the mathematical formulation and the properties of RBO are described. Afterwards, the original definition of the single RBO score (RBO$_{\text{EXT}}$) is provided, alongside an analysis of the assumptions and limitations behind this point estimate.

2.1 Properties of Rank-Biased Overlap

In order to compare rankings and infer details about the algorithms that produced them, similarity measures can be calculated as a substitute for the highly subjective relevance judgements of human reviewers [Bar-Ilan *et al.*, 2006]. However, in practice, rankings are typically indefinite (meaning that any truncation depth can be selected for evaluation), incomplete (resulting in non-conjointness as the two rankings only cover subsets of the entire domain), and top-weighted (making the ordering at the top of the two lists more important than that further down the tails) [Webber *et al.*, 2010].

Introduced in 2010, RBO is a similarity measure designed to accommodate rankings with those three properties. To handle incompleteness, RBO is based on overlap (i.e. set intersection) rather than correlation [Webber *et al.*, 2010]. Specifically, overlap is defined as the number of items that the two indefinite rankings S (with a visible prefix of length s) and L (with its typically longer visible prefix of length $l \geq s$) have in-common up to evaluation depth d:

$$X_{S,L,d} = |S_{:d} \cap L_{:d}| \tag{1}$$

The proportion of overlapped items at d is then referred to as agreement:

$$A_{S,L,d} = \frac{X_{S,L,d}}{d} \tag{2}$$

To introduce top-weightedness, RBO is parameterized by the persistence value p $(0 < p < 1)$, which denotes the probability of a user continuing to consider the items ranked next in the two lists [Moffat and Zobel, 2008]. In turn, the probability of the evaluation terminating at the current depth is $1 - p$.

For RBO to handle indefinite rankings and output a consistent similarity score regardless of the chosen truncation depth, fixed weights are assigned to each rank following the convergent geometric series, whose infinite sum is defined as:

$$\sum_{d=1}^{\infty} p^{d-1} = \frac{1}{1-p} \tag{3}$$

To sum up to 1, each weight is modeled as $w_d = (1-p) \cdot p^{d-1}$ [Webber *et al.*, 2010]. As a consequence, the weights are also decreasing, which causes the visible prefixes to contribute to the similarity score more than the infinite tails [Clarke *et al.*, 2020]. RBO can then be derived as:

$$RBO(S, L, p) = (1-p) \sum_{d=1}^{\infty} A_d \cdot p^{d-1} \tag{4}$$

The output of RBO can range from 0 to 1, where 0 indicates that the two rankings are completely dissimilar (i.e. fully disjoint), and 1 implies they contain the same elements [Webber *et al.*, 2010]. In essence, based solely on an evaluation of the finite visible sections of the two rankings, RBO imposes tight bounds on the full similarity score [Webber *et al.*, 2010]. As the evaluation depth increases, both bounds become narrower, and the approximation's accuracy improves.

RBO can therefore be applied in many scenarios where indefinite rankings are generated, and other similarity measures inspired by RBO have been developed to satisfy various new properties and make alternative trade-offs [Cardoso and Magalhães, 2011; Tan and Clarke, 2014]. New variants of RBO were also recently proposed that account for the presence of ties in the visible prefixes, extending the original framework from 2010 [Corsi and Urbano, 2024].

2.2 Assumptions and Limitations of Extrapolated Rank-Biased Overlap

In most cases where RBO is used, a single similarity score is often reported and analyzed. To find the latter, the agreement seen at the last depth of the visible prefixes is assumed to continue indefinitely across the unseen parts of the two rankings [Webber *et al.*, 2010]. Thus, the point estimate RBO$_{\text{EXT}}$ is extrapolated out from this agreement according to the following formula for visible-prefix lengths $l \geq s$:

$$\begin{aligned} RBO_{EXT}(S, L, s, l, p) = \frac{1-p}{p} \Bigg(\sum_{d=1}^{l} \frac{X_d}{d} p^d + \\ + \sum_{d=s+1}^{l} \frac{(d-s)X_s}{ds} p^d \Bigg) + \left(\frac{X_l - X_s}{l} + \frac{X_s}{s} \right) p^l \end{aligned} \tag{5}$$

If the visible prefixes are of the same length, the procedure is to just compute the weighted sum of all observed agreements from 1 to k $(= s = l)$ and after that extrapolate the agreement at the last depth A_k.

However, when $l > s$, a separate extrapolation occurs for the $l - s$ items that are unseen in S (i.e. the section $S_{(s+1):l}$). As shown in Equation 5, those elements are assigned a probability of membership equal to A_s, the agreement observed at depth s [Webber *et al.*, 2010].

Based on this assumption, the assumed agreement at depth l ($\tilde{A}_l$) is then computed and extrapolated for items in the unseen tails of the two rankings (the last term of Equation 5):

$$\tilde{A}_l = \frac{X_l + (l-s) \cdot A_s}{l} = \frac{X_l - X_s}{l} + \frac{X_s}{s} \tag{6}$$

The assumption of constant agreement that the two extrapolations (of A_s and $\tilde{A}_l$) are based on, however, does not account for any patterns occurring in the visible prefixes, limiting the available information regarding agreement-behavior to solely those two snapshots at depths s and l. This inflexibility could result in a divergence between assumed and real agreement if for example, agreement remains low in the visible parts, only to then asymptotically approach 1 throughout the unseen tails. Such a divergence could in turn heavily lower the accuracy of RBO$_{\text{EXT}}$, especially when the seen section is short while p is close to 1, causing a lot of unseen items to remain significant for the RBO score due to their large weights.

Considering the potentially detrimental effects of such inflexibility on the accuracy of RBO$_{\text{EXT}}$, the aim of the study is to redefine the single RBO score under more relaxed assumptions in an effort to achieve more accurate approximations of both agreement and RBO. Thus, a new mathematical framework for RBO$_{\text{EXT}}$ is the focus of the next section.

3 Proposed Reformulations of Agreement and Extrapolated Rank-Biased Overlap

Altering the assumption of constant agreement for the unseen tails of the two rankings S and L is the key to determining if the accuracy of the single RBO score can be improved. Thus, a redefinition of RBO$_{\text{EXT}}$ and new methods for computing the assumed agreement at any depth d ($\tilde{A}_d$) constitute the study's main contributions and are outlined in this section.

As done in previous redefinitions of RBO, the formula for the similarity score can be split into three sections with regard to the scope of the visible prefixes [Corsi and Urbano, 2024]. These three intervals contain the items visible in both prefixes (depths 1 to s), the remainder of the longer prefix with a corresponding unseen section of S (depths $s+1$ to l), as well as the elements that are unseen in both rankings (depths $l+1$ to ∞), respectively. Such a separation results in the redefinition of the RBO point estimate below:

$$RBO_{EXT}^{REDEF}(S, L, s, l, p) = \frac{1-p}{p}\Bigg(\sum_{d=1}^{s} A_d p^d + \\ + \sum_{d=s+1}^{l} \tilde{A}_d p^d + \sum_{d=l+1}^{\infty} \tilde{A}_d p^d\Bigg) \tag{7}$$

In the first section, A_d is simply the observed agreement since the items in both prefixes are seen: it is thus calculated using Equation 2. Meanwhile, the assumed agreement from the last two terms of Equation 7 is formulated differently with respect to the interval that d falls within:

$$\tilde{A}_d = \begin{cases} \dfrac{X_d \,+\, \sum_{k=s+1}^{d} \hat{A}_k}{d} & d \in [s+1; l] \\ \hat{A}_d & d \in [l+1; \infty) \end{cases} \tag{8}$$

When items in one or both rankings are unseen, it is necessary to assign those elements a degree of membership – or in other words, their estimated contribution to the assumed agreement at the chosen depth d. This is precisely the role of $\hat{A}_k$ (or $\hat{A}_d$ for the second part of Equation 8), which is interpreted as the probability that an element selected at random appears in both rankings [Webber *et al.*, 2010]. Estimating this probability of membership and assigning it to the unseen items results in the two formulations of assumed agreement (Equation 8).

For the $[s+1; l]$ interval, the observed overlap X_d (which is computed using Equation 1) accounts for the $l - s$ remaining seen items in the longer prefix. The extrapolation is therefore partial, the contribution of the unseen elements $S_{(s+1):l}$ set to $\hat{A}_k$. In contrast, beyond l, both S and L are unknown, and no further overlap is inferred: thus, $\tilde{A}_d$ is simply equal to $\hat{A}_d$.

With Equation 8, the assumed agreement can be calculated at any depth from $s+1$ to ∞ without extrapolating the values of A_s and $\tilde{A}_l$ as done in the original framework and explained in Section 2.2. Instead of those two constant agreements, the degree of membership $\hat{A}_k$ is computed at every depth beyond s, in an effort to introduce variability into the assumed agreement and, by extension, the extrapolated similarity score.

To that end, this study offers three approaches to calculate an unseen item's estimated contribution $\hat{A}_k$ (outlined below):

Previous-Value Approach

This implementation simply re-uses the value of the assumed agreement at the previous depth as the estimated contribution of the current element, namely $\hat{A}_k = \tilde{A}_{k-1}$.

Logistic-Regression Approach

In this formulation, the value for $\hat{A}_k$ is assigned as the output of logistic regression. The model computes a linear combination of the independent variable (in this case, depth), feeding the result to the standard logistic function:

$$\hat{A}_k = \frac{1}{1 + e^{-(\beta_0 \,+\, \beta_1 \cdot k)}} \tag{9}$$

The sigmoid is used to impose a range of $[0; 1]$ on the output for $\hat{A}_k$. Meanwhile, the intercept (β_0) and the coefficient (β_1) are optimized via a regression-fit on the observed agreements from depths 1 to s.

Generalized-Additive-Model Approach

The third approach uses a generalized additive model (GAM) as the more flexible alternative to ordinary logistic regression and its underlying linear combination [Hastie and Tibshirani, 1986]. $\hat{A}_k$ is calculated as follows:

$$\hat{A}_k = \frac{1}{1 + e^{-[s_0 \,+\, s_1(k)]}} \tag{10}$$

The main advantage of GAMs allowing them to more closely capture patterns in the data they are fitted against is the use of non-parametric smoother functions, which may be non-linear with respect to the independent variable (here, the evaluation depth k) [Hastie and Tibshirani, 1986]. Therefore, $\beta_1 \cdot k$ from Equation 9 is replaced by $s_1(k)$, with a penalized cubic spline chosen as the function's basis [Hastie and Tibshirani, 1986].

The sigmoid and the intercept term (s_0) are otherwise identical in terms of the role they fulfill in Equations 9 and 10. As the sigmoid is applied here, too, this approach is also referred to as Logistic-GAM throughout the paper.

4 Experimental Setup and Results

This section demonstrates the main components of the evaluation procedure followed throughout the study, offering more details on the use of synthetic data, the setup of different testing configurations, and the choice of a performance measure. Additionally, it provides an overview of the results as well as a comparison between the accuracy achieved by the proposed RBO$_{\text{EXT}}$ redefinitions and that of the original formulation. In this manner, the empirical analysis is based on the theoretical framework and the assumptions outlined in Sections 2 and 3.

4.1 Usage of Simulated Rankings

As discussed in Section 2.1, most rankings are incomplete as they do not contain every possible item from the domain. Due to this, two given rankings are in-practice likely non-conjoint, each covering a different subset of the full domain.

To control for the degree of incompleteness and (by extension) non-conjointness, an algorithm for generating synthetic rankings was used in which the number of unique items in the rankings' shared domain is a freely-tunable parameter. A link to the data-generation code, which was written by the authors of several tie-handling RBO variants, can be found in Section 7 of the paper [Corsi and Urbano, 2024].

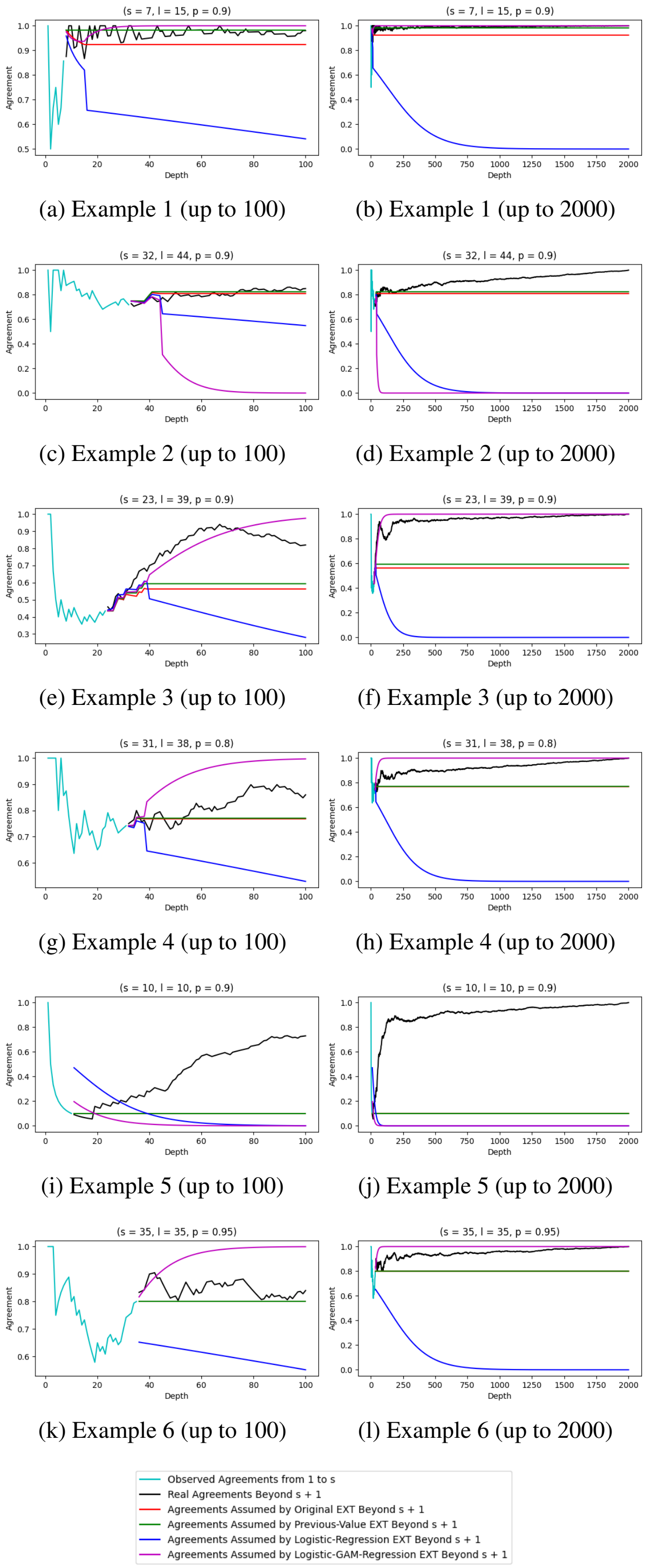
The domain size chosen for this study was 2000, and 5000 pairs of rankings were generated, each ranking of length 2000 elements. Having the domain and the full-ranking sizes equal produced fully-conjoint rankings whose agreement reaches 1 at depth 2000. To control for the degree of non-conjointness, therefore, the values for s and l could be varied, yielding two visible prefixes that satisfy the property of incompleteness by only covering (different) subsets of the whole domain.

Three values were selected for the persistence parameter p: 0.8, 0.9, and 0.95. The number of expected observed items is thus 5, 10, or 20 – based on $\omega = \frac{1}{1-p}$ [Webber *et al.*, 2010].

Finally, the lengths of the two visible prefixes s and l were chosen using a pseudo-random number generator with respect to the top-weightedness (ω) and the length of the full rankings (∞ in practice – 2000 for this study, as explained earlier). To leave room for the unseen tails and introduce variations in the amount of seen items, l and s were generated as follows:

$$\begin{aligned} l &= randint(\omega,\ 45) \\ s &= randint(\lfloor 0.75 \cdot \omega \rfloor,\ l) \end{aligned} \tag{11}$$

In this manner, the majority of the weight-significant items in the two lists were guaranteed to fall within the observed $[1; s]$ section, and even rankings ($s = l$) could also be generated.



(a) Example 1 (up to 100) (b) Example 1 (up to 2000)

(c) Example 2 (up to 100) (d) Example 2 (up to 2000)

(e) Example 3 (up to 100) (f) Example 3 (up to 2000)

(g) Example 4 (up to 100) (h) Example 4 (up to 2000)

(i) Example 5 (up to 100) (j) Example 5 (up to 2000)

(k) Example 6 (up to 100) (l) Example 6 (up to 2000)

Figure 1: Actual and assumed agreements in six selected scenarios. The legend indicates the range and type of each agreement-trace; the s / l / p configurations for every scenario are listed above the plots.

4.2 Evaluation Procedure

In order to conduct a fine-grained evaluation where the different factors influencing the results are easy to trace and reason about, the randomly-generated lengths s were split into three categories: small ($s \leq 15$), medium ($15 < s \leq 30$), and large ($s > 30$). In combination with the 3 values chosen for p, this resulted in 9 distinct configurations in which the performance of the four RBO_{EXT} formulations could be compared.

To conduct this assessment, two criteria were selected (and applied equally to all four RBO_{EXT} definitions): the accuracy of the RBO point estimate and the accuracy of the agreement assumptions in the presence of unseen items – namely, for all depths in the interval $[s + 1; \infty)$.

In turn, the performance measure of *accuracy* was defined as closeness to the real values for RBO and agreement. In the case of RBO-accuracy, the distance (i.e. absolute difference) between the given point estimate and the true RBO similarity score was calculated. As a value for agreement-accuracy, the average distance between the assumed and the real agreement quantities at depths $s + 1$ to ∞ (i.e. 2000) was computed.

Lastly, it is significant to illustrate how the real RBO score and the actual agreements beyond s were derived. The former was treated as the output of RBO_{EXT} when given the two full lists S and L up to depth 2000 (in other words, exhaustively computing the agreements to ∞). This arrives precisely at the true similarity score and also provides the real agreements in the interval $[s+1; \infty)$ – both necessary to calculate accuracy.

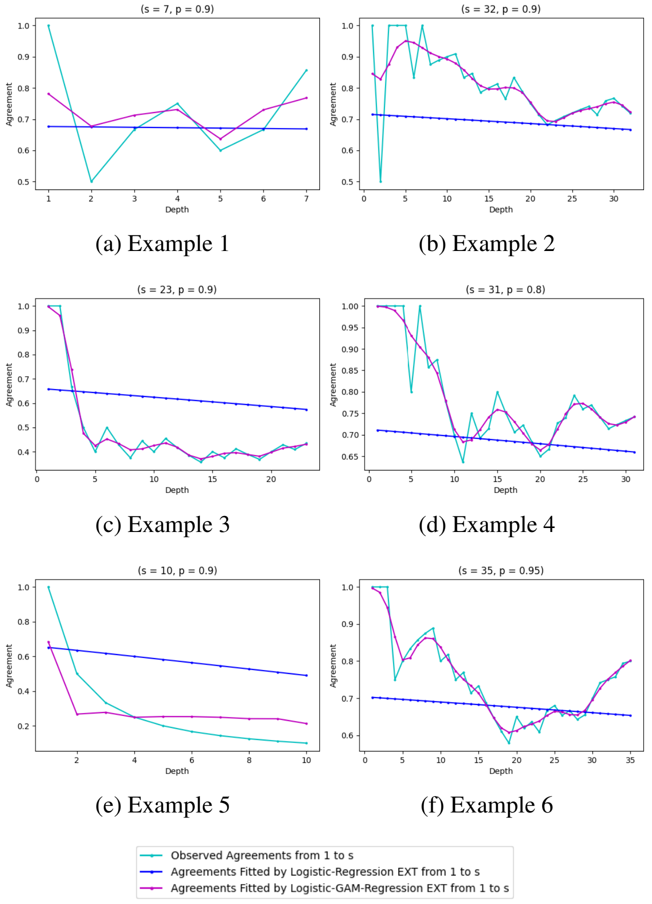
4.3 Results and Observations

The results aggregated throughout this experiment constitute measurements of the RBO- and agreement-accuracy achieved by the RBO_{EXT} implementations. These performance criteria are separately analyzed in the following two sub-sections.

RBO-Accuracy

Following the outlined evaluation procedure, the agreement- and RBO-accuracy of all four RBO_{EXT} implementations was measured in different configurations with respect to p and the type of randomly-generated s. Tables 1 and 2 provide aggregated information about RBO-accuracy for fixed values of p (0.8 and 0.95, respectively). Four measures are presented for each category of RBO_{EXT} and s: the average RBO-accuracy, the maximal observed absolute difference, as well as the percentages of RBO-distances that qualify as medium (between 0.01 and 0.1) and large (greater than 0.1).

For both values of p and across all three intervals for s, the original RBO similarity score is the most accurate on average, and the point estimate computed using the logistic-regression approach performs the poorest. The average RBO-distances observed for the latter are 7 to 23 times those for the original implementation: for example, 0.0517 vs. 0.0076 (Table 1) or 0.0961 compared to 0.0070 (Table 2). The logistic-regression approach is also characterized by the highest percentages of medium and large RBO-distances (e.g. 62% and 18% or 25% and 73% in the category $s \leq 15$). The reason for these large absolute differences is the inflexible logistic regression model underfitting the observed agreements at depths 1 to s, which results in inaccurate predictions for $\hat{A}_k$, a divergence between assumed and actual agreements beyond depth $s + 1$, and thus (on average) an inaccurate RBO point estimate.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

(e) Example 5

(f) Example 6

Figure 2: Closeness-of-fit for Logit-Regression and Logistic-GAM on the observed agreements up to s (from the scenarios in Figure 1). The legend indicates the range and type of each agreement-trace; the s / p configurations for every scenario are listed above the plots.

In contrast to simple logistic regression, all average RBO-distances for the Logistic-GAM formulation are substantially smaller, and the percentages of large differences are reduced (e.g. 21% and 12% compared to 73% and 49% in Table 2). It is worth noting, however, that Logistic-GAM has much larger maximal RBO-distances than all other RBO redefinitions for $p = 0.95$ (Table 2), which implies that greater flexibility does not yield a better regression-fit in every scenario. Meanwhile, the performance of the previous-value approach in all settings for p and s closely resembles that of the original extrapolated score – in terms of average, maximum, and M | L percentages.

Broadly inspecting the two tables, the effects of s and p on overall RBO-accuracy also become apparent. As the length s transitions from small to large, the average RBO-distance for all four RBO_{EXT} formulations decreases. The more items are present in the fully-visible section (up to depth s), the smaller the absolute differences become, dropping below 0.1 (flowing from L into M) or below 0.01 (falling out of M entirely). This can be observed in the last two columns of Table 2: for GAM,

Table 1: Summarized measures of RBO-accuracy for a fixed $p = 0.8$ across the three categories of s (small, medium, and large). M stands for medium RBO-distances in the interval $(0.01; 0.1]$, and L represents large RBO-distances in the range $(0.1; 1]$.

Type of s	$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{OG}}\rvert$				$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{PV}}\rvert$				$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{LR}}\rvert$				$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{GAM}}\rvert$			
	Avg.	**Max.**	**M**	**L**	**Avg.**	**Max.**	**M**	**L**	**Avg.**	**Max.**	**M**	**L**	**Avg.**	**Max.**	**M**	**L**
$s \leq 15$	0.0076	0.2173	17%	1%	0.0077	0.1974	17%	1%	0.0517	0.2567	62%	18%	0.0116	0.2404	25%	2%
$15 < s \leq 30$	0.0001	0.0029	0%	0%	0.0001	0.0029	0%	0%	0.0017	0.0141	1%	0%	0.0003	0.0104	0%	0%
$s > 30$	$3.70 \cdot 10^{-6}$	$5.3 \cdot 10^{-5}$	0%	0%	$3.76 \cdot 10^{-6}$	$5.6 \cdot 10^{-5}$	0%	0%	$8.53 \cdot 10^{-5}$	$5.0 \cdot 10^{-4}$	0%	0%	$9.99 \cdot 10^{-6}$	$3.4 \cdot 10^{-4}$	0%	0%

Table 2: Summarized measures of RBO-accuracy for a fixed $p = 0.95$ across the three categories of s (small, medium, and large). M stands for medium RBO-distances in the interval $(0.01; 0.1]$, and L represents large RBO-distances in the range $(0.1; 1]$.

Type of s	$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{OG}}\rvert$				$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{PV}}\rvert$				$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{LR}}\rvert$				$\lvert\mathbf{RBO} - \mathbf{RBO}_{\mathbf{EXT}}^{\mathbf{GAM}}\rvert$			
	Avg.	**Max.**	**M**	**L**	**Avg.**	**Max.**	**M**	**L**	**Avg.**	**Max.**	**M**	**L**	**Avg.**	**Max.**	**M**	**L**
$s \leq 15$	0.0116	0.1127	35%	0%	0.0127	0.1154	36%	0%	0.1231	0.1931	25%	73%	0.0763	0.3273	77%	21%
$15 < s \leq 30$	0.0070	0.0930	25%	0%	0.0075	0.0930	26%	0%	0.0961	0.1969	46%	49%	0.0378	0.3164	40%	12%
$s > 30$	0.0025	0.0334	6%	0%	0.0025	0.0330	6%	0%	0.0523	0.1020	92%	0%	0.0124	0.1612	32%	1%

for instance, M + L equals 98%, 52%, and 33% (with an ever-smaller share assigned to L) as s increases. In turn, the larger value $p = 0.95$ causes RBO to be less top-weighted, resulting in the noticeably greater RBO-distances presented in Table 2. More of the assumed agreements beyond depth s, which may drastically diverge from the actual values, become significant during the computation of RBO_{EXT}, decreasing its accuracy.

Agreement-Accuracy

To demonstrate how assumed agreements beyond depth s are calculated using the four extrapolation approaches, 6 specific scenarios were selected from the aggregated results such that the various assumptions' effects are apparent. Figure 1 shows the assumed agreements for each scenario twice: the first plot is truncated at depth 100 to clearly visualize trends in the most significant agreements in terms of weight, whereas the second one continues to the maximal depth of 2000. Figure 2 in turn presents how well the two regression-based approaches fit on the observed agreements (depths 1 to s), which fulfill the role of training data for the models.

One difference that is immediately clear is the non-linearity of the agreement-traces for both regression-based approaches in contrast to the eventually-constant agreements assumed by the original and previous-value extrapolations. Capturing the patterns in the observed agreements via a regression-fit which tunes the parameters of Equations 9 and 10, GAM and logistic regression have greater flexibility, which in the case of GAM can significantly outperform other approaches with respect to agreement-accuracy (Examples 1 and 3). Moreover, accurate estimations of $\hat{A}_k$ result in a noticeably higher RBO-accuracy when the agreement at depth k is significant: this is illustrated by GAM's small RBO-distances in Ex. 1 and 3 (Table 3).

In other scenarios, however, that same flexibility can cause low agreement-accuracy compared to simpler extrapolations. Examples 2 and 5 show that if the trends captured throughout depths 1 to s do not persist in the unseen remainder of the two rankings, logistic regression and GAM produce estimates that drastically diverge from the real agreement-values (Figure 1).

Such mismatch is only exacerbated by GAM's tendency to overfit on the observed agreements as indicated in all plots of Figure 2. Compared to logistic regression, GAM's far steeper trace of fitted agreements in Ex. 2 and 5 leads to the very low agreement-accuracy beyond s (Figure 1). Underfitting, as the opposite extreme, characterizes logistic regression, and it can have an equally detrimental effect on agreement-accuracy. In Examples 1, 4, and 6, the closely-fitting GAM captures trends of increase towards depth s (Figure 2), allowing it to perform better than the inflexible logistic regression (Figure 1). When the visible prefixes are even (Example 6), accounting for such patterns during training becomes especially crucial as there is no section $[s + 1; l)$ and thus no further overlap to consider.

Lastly, it is worth pointing out that low agreement-accuracy does not necessarily imply poor RBO_{EXT} performance. GAM and logistic regression both output inaccurate assumed agreements in Example 2 (Figure 1), yet the RBO-distances shown in Table 3 do not differ drastically among the four approaches (GAM still performing the worst with a difference of 0.0057). These results confirm that agreements further from the visible prefixes are less significant for computing the RBO score, no matter how well they approximate their real counterparts.

5 Discussion and Limitations

Linking back to the properties of rankings outlined in Section 2 and the mathematical foundations of the proposed RBO_{EXT} reformulations from Section 3, this section places the study's findings in a broader context and reflects on the generalizability of the results. Therefore, the limitations of the experiment are established as an essential factor when analyzing how the altered agreement-assumptions affect the accuracy of RBO.

While it remains important that the agreements assumed in the various extrapolation approaches are overall accurate with respect to the real values, the range of agreements significant for the final RBO value is determined by the value of p. When measuring the performance of RBO_{EXT}, the decaying weights w_d and the resulting degree of top-weightedness described in Section 2.1 need to be accounted for as a smaller p causes the

Table 3: Measures of RBO-accuracy in the six scenarios from Figures 1 and 2.

RBO-Accuracy	**Example 1**	**Example 2**	**Example 3**	**Example 4**	**Example 5**	**Example 6**
$\lvert RBO - RBO_{EXT}^{OG}\rvert$	0.0063	0.0009	0.0067	$7.50 \cdot 10^{-6}$	0.0101	0.0091
$\lvert RBO - RBO_{EXT}^{PV}\rvert$	0.0138	0.0012	0.0057	$6.75 \cdot 10^{-6}$	0.0101	0.0091
$\lvert RBO - RBO_{EXT}^{LR}\rvert$	0.0811	0.0010	0.0065	$4.05 \cdot 10^{-5}$	0.0714	0.0385
$\lvert RBO - RBO_{EXT}^{GAM}\rvert$	0.0048	0.0057	0.0030	$1.48 \cdot 10^{-5}$	0.0055	0.0121

unseen tails to matter very little (save for the first few ranks if the length s is also small, for instance). Therefore, even if the assumed agreements completely diverge from reality towards later depths, the RBO score would largely be unaffected.

As for the visible-prefix lengths s and l, they are key factors in determining the accuracy of RBO_{EXT} as the visible parts of the two rankings carry the largest weight. That is why a larger value of s (i.e. access to more of the actual agreements) leads to a smaller RBO-distance for all four RBO_{EXT} formulations. The section $[s+1; l)$ is similarly beneficial for accuracy since the seen items in L still contribute to overlap (the X_d term in the first part of Equation 8). If l is far greater than s, however, this partial extrapolation would dominate the two fully unseen tails (beyond l) during the calculation of RBO and inflate the final score, preventing the effects of the modified assumption regarding constant agreement from becoming apparent. In an effort to prevent this and make the interval $[l+1; \infty)$ equally represented during experimentation, the upper threshold for l in random number generation was set to 45 (Equation 11).

Accounting for the patterns of agreement up to depth s lies at the center of the study's efforts to redefine RBO_{EXT}, and it is the primary criterion based on which the four formulations can be differentiated. The simpler original and previous-value approaches consider limited information during extrapolation ($A_s|\tilde{A}_l$ and $\tilde{A}_{d-1}$, respectively), and the assumed agreements they produce remain constant beyond the visible prefixes. As a more flexible alternative, the two regression-based RBO_{EXT} implementations involve fitting a model on the agreements up to s that then outputs estimates of the membership probability $\hat{A}_k$ at every depth to ∞. The use of regression introduces the trade-off between under- and overfitting in the training phase, which is precisely the distinction between logistic regression and GAM. While the latter better reproduces the trends in the seen agreements and has a higher average RBO-accuracy, the former's tendency to underfit enables it to be less impacted if the patterns of observed agreement happen to be inconsistent with the actual agreement-values further along the tails.

Despite offering insights into the strong points, drawbacks, and performance of the extrapolation techniques evaluated in this study, the results presented in Section 4.3 of the paper do not generalize well to realistic scenarios in which RBO might typically be applied. This is caused by the fully-conjoint simulated pairs of rankings whose agreement is guaranteed to be 1 at the maximal depth – something that in-practice is uncharacteristic of incomplete and indefinite rankings. Therefore, to preserve the property of incompleteness and ensure that S and L remain non-conjoint within the visible prefixes, the domain size (2000) was chosen to be far greater than the upper bound for l (45). Despite this workaround, the actual agreements for the unseen tails remain unrealistically large, discouraging the generation of longer visible prefixes for the experiment while also imposing an inconveniently-high baseline-agreement for the measurement of agreement-accuracy. Ultimately, the tails being poorly-generalizable was the reason why p was chosen further from 1 (0.8 to 0.95), increasing top-weightedness and preventing RBO-accuracy from being undesirably distorted.

6 Conclusions and Future Work

In this paper, the RBO point estimate as well as its assumption of constant agreement throughout the unseen tails of any two rankings were critically assessed. In the search of some more accurate extrapolation method, three RBO_{EXT} reformulations with relaxed assumptions about agreement in the unseen parts were proposed, and their closeness to the real RBO score was measured for different values of p and the prefix-lengths $s|l$.

Through the devised experiment, it was discovered that the original and previous-value extrapolations perform the best in terms of RBO-accuracy, despite their assumed agreements in the unseen tails remaining constant up to the maximum depth. The remaining two RBO_{EXT} redefinitions utilized regression, fitting a function on the observed agreements to depth s with the purpose of more closely capturing the patterns in that first section. The simpler logistic-regression approach exhibited a tendency of underfitting, which resulted in its RBO-accuracy being the lowest overall. In contrast, the generalized additive logistic model (Logistic-GAM) far better replicated the trends in the seen agreement-values, achieving higher accuracy than logistic regression yet becoming susceptible to any mismatch between the patterns in the visible prefixes and in the tails.

There were, however, limitations imposed on the study that are important to consider when interpreting its results. As the simulated rankings used for the experiment had an unrealistic agreement close to 1 across later depths in the unseen section, only relatively short visible prefixes of S and L and values of p further from 1 were considered for the evaluation. This was necessary in order to avoid the RBO-accuracy measurements becoming skewed due to the poorly-generalizable agreements in the rankings' unseen parts.

Thus, some appropriate directions for future work include:

- Using simulation code that is better suited towards RBO and its properties (generating more realistic, incomplete rankings whose agreement does not tend to 1 at ∞);

- With such more appropriate rankings, investigating both greater values for p and longer "indefinite" lists S and L (as the rankings would no longer be fully-conjoint, their tails would generalize well and be informative, meaning that the degree of top-weightedness could be reduced);
- Reformulating RBO_{EXT} under relaxed assumptions with respect to alternative implementations of RBO – such as its three tie-handling variants [Corsi and Urbano, 2024].

7 Responsible Research

This section outlines the precautions and generalizability concerns relevant for this study, with the purpose of maintaining transparency and preserving academic integrity. Additionally, it provides an overview of the steps taken to uphold the reproducibility of the chosen methods.

First, it is important to emphasize that the newly-proposed RBO_{EXT} formulations are based on entirely different assumptions compared to the original implementation and might thus no longer be suitable for all scenarios in which RBO is being applied currently. Therefore, prior to using any of these three redefinitions, readers should consider the modified theoretical framework and identify whether the given RBO_{EXT} variant is still appropriate for their use-case.

Additionally, as highlighted in Section 5, the data used for the study were simulated pairs of fully-conjoint rankings with an uncharacteristically-high agreement in the unseen tails. As a consequence, the agreement-accuracy measured throughout the experiment was based off an unrealistic baseline of actual agreements approaching 1 beyond the visible prefixes. As for RBO-accuracy, shorter lengths l and smaller values of p were investigated to prevent the ungeneralizable tails from skewing the computed absolute differences. These limitations need to be kept in-mind when referring to the results of this study.

The simulation code used to generate the data is published online,[1] the authors having used it to evaluate their own RBO redefinition [Corsi and Urbano, 2024]. The full data-files that contain the simulated pairs of rankings can be freely-accessed and reused from the public GitHub repository created for this experiment, also available online.[2]

This repository also includes all results (JSON format) and figures (PNG format) that were aggregated from the 9 combinations of p and the category of s (refer to Section 4.2). Only a subset of those measurements was presented in the paper as many of them are summarized by the averages from Tables 1 and 2. Due to space considerations, only the truly-informative instances were provided in Figures 1 and 2 since the RBO_{EXT} formulations and their assumptions are well-expressed there.

The source code required to run the experiment is provided as well, alongside a file listing the required modules: *numpy*[3] (used for computing the average accuracy scores), *matplotlib*[4] (used for producing plots), and *pygam*[5] (an implementation of GAMs in Python). Furthermore, a detailed README file has been included, with all default values for the hyperparameters specified. For example, the seed used for the pseudo-random number generation of s and l has a fixed value of 42 – if kept the same, it allows for the results to be reproduced.

Finally, citations to other works have been provided for all borrowed ideas throughout the paper. Readers could this way refer to those contributions for additional details.

[1] https://github.com/julian-urbano/sigir2024-rbo

[2] https://github.com/Konstantin-Asen/cse3000-research-project

[3] https://pypi.org/project/numpy/

[4] https://pypi.org/project/matplotlib/

[5] https://pypi.org/project/pygam/

References

[Bar-Ilan *et al.*, 2006] Judit Bar-Ilan, Mazlita Mat-Hassan, and Mark Levene. Methods for comparing rankings of search engine results. *Computer Networks*, 50(10):1448–1463, 2006.

[Bar-Ilan, 2005] Judit Bar-Ilan. Comparing rankings of search results on the Web. *Information Processing & Management*, 41(6):1511–1519, 2005.

[Cardoso and Magalhães, 2011] Bruno Cardoso and João Magalhães. Google, Bing and a New Perspective on Ranking Similarity. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1933–1936, New York, NY, USA, 2011. Association for Computing Machinery.

[Clarke *et al.*, 2020] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 225–234, New York, NY, USA, 2020. Association for Computing Machinery.

[Corsi and Urbano, 2024] Matteo Corsi and Julián Urbano. The Treatment of Ties in Rank-Biased Overlap. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[Fagin *et al.*, 2003] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[Hastie and Tibshirani, 1986] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297–318, 1986.

[Moffat and Zobel, 2008] Alistair Moffat and Justin Zobel. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.*, 27(1), 12 2008.

[Tan and Clarke, 2014] Luchen Tan and Charles L. A. Clarke. A Family of Rank Similarity Measures Based on Maximized Effectiveness Difference. *IEEE Transactions on Knowledge and Data Engineering*, 27:2865–2877, 2014.

[Webber *et al.*, 2010] William Webber, Alistair Moffat, and Justin Zobel. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.*, 28(4), 11 2010.