Extending Rank-Biased Overlap (RBO) to Relevance Profiles

Author: Thijs Houben (t.h.j.houben@student.tudelft.nl) [1] Supervisor & Responsible Professor : Julián Urbano (j.urbano@tudelft.nl) [1]

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

1 - Introduction

- Documents in rankings can be assigned a **graded relevance** score from 0 to M to indicate how **relevant** they are to a user.
- Traditional rank similarity metrics like **Rank-Biased Overlap (RBO)** [3] compare rankings based on **item identity** only.
- This **ignores** cases where different items may have equal or similar relevance to a user:
  - Ranking S: ⟨doc1 (rel=3), doc2 (rel=2), doc3 (rel=0)⟩
  - Ranking L: ⟨doc4 (rel=3), doc5 (rel=2), doc6 (rel=0)⟩
- RBO would state that these are **completely dissimilar**.
- Therefore, similarity based on **relevance**—rather than just **item overlap**—is needed.

Research Question: *How can Rank-Biased Overlap (RBO) be extended for relevance values?*

2 - Background: Identity-Based RBO

- **RBO** is a rank similarity metric that:
  - Is **top-weighted**, giving more weight to agreements near the top of the ranking.
  - Supports **non-conjoint** lists, rankings do not need to have the same items.
  - Can handle ranking pairs L and S of **different** sizes ($|S| = s \leq l = |L|$).
  - Assumes the rankings provided are **prefixes** of **indefinite** rankings.
- RBO calculates **agreement** between rankings S^{id} and L^{id} at each rank depth d as the proportion of **overlapping** items up to that depth:

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

- RBO is a **weighted sum** of these agreements:

$$\text{RBO}_{L,S,p} = \frac{1-p}{p}\left(\underbrace{\sum_{d=1}^{s} A_{L,S,d} \cdot p^d}_{1} + \underbrace{\sum_{d=s+1}^{l} A_{L,S,d} \cdot p^d}_{2} + \underbrace{\sum_{d=l+1}^{\infty} A_{L,S,d} \cdot p^d}_{3}\right)$$

  where $p \in (0, 1)$ is a **persistence** parameter controlling how **top-weighted** the metric is.
- The RBO equation is split up in 3 parts, where at depth d:
  - Part 1: Items of both L and S are known.
  - Part 2: Item of S is unknown, L is known.
  - Part 3: Items of both rankings are unknown.
- Three scores are calculated, with **different assumptions** about the **unknown** parts:
  - RBO_{MIN} assumes the **worst-case** continuation for the unseen parts.
  - RBO_{MAX} assumes the **best-case** continuation for the unseen parts.
  - RBO_{EXT} **extrapolates** overlap and agreement based on the prefixes.

3 - Relevance-Based RBO

- Agreement is redefined in terms of **cumulative gain (CG)** [2] to compare rankings based on their **relevance profiles**.
- **Gain** quantifies how much utility a document with relevance score x provides. Two common gain functions, with the adjustable hyperparameter θ, are:
  - Linear: $G_x = \theta \cdot x$,
  - Exponential: $G_x = \theta^x - 1$.
- **CG** is the sum of **gain scores** up to depth d in a ranking.
- **Relevance-based agreement** at depth d is defined as:

$$A^{\text{CG}}_d = 1 - \frac{|\text{CG}_{S,d} - \text{CG}_{L,d}|}{\mathcal{N}_d}$$

  where $\mathcal{N}_d$ is a **normalization** factor.
- Plugging this A_d into the RBO formula yields a new metric reflecting the **informational similarity** of the rankings, rather than just **overlap**.

4 - Different normalization factors

Two normalization approaches are proposed:

- **Local:** $\mathcal{N}^{\text{loc}}_d = \max\{CG_{L^{\text{rel}},d}, CG_{S^{\text{rel}},d}\}$ normalizes by the **maximum observed** cumulative gain at depth d. (Edge cases due to CG = 0 have been omitted).
- **Global:** $\mathcal{N}^{\text{glo}}_d = G_M \cdot d$ normalizes by the **maximum possible** cumulative gain at depth d, where M is the maximum relevance score.

5 - Calculating RBO_{MIN}, RBO_{MAX} and RBO_{EXT}

- 1. RBO_{MIN} - Worst-case continuation
  - Part 2: Construct a **dynamic programming table** with the **minimum** RBO score for all reachable CG at depth d. Then, select score from table that together with part 3 leads to the minimum score.
  - Part 3: Assume difference **increases** by G_M at each depth and calculate using closed form equations.
- 2. RBO_{MAX} — Best-case continuation
  - **No** general method was found to calculate RBO_{MAX} for all possible gain functions, just for **linear**.
  - Part 2: **Greedily** assume continuation of S^{rel} that **minimizes** the difference between CG scores.
  - Part 3: Assume difference **decreases** by G_M at each depth and calculate using closed form equations.
- 3. RBO_{EXT} — Extrapolated continuation
  - Part 2: $CG_{S^{\text{rel}},d} = \frac{d \cdot CG_{S^{\text{rel}},s}}{s}$, **extrapolate** based on the cumulative gain of the **prefix** of S^{rel}.
  - Part 3: $A_d = A_l$, **assume** agreement stays the same and calculate using closed form equation.

6 - Test setup and Results

- Tests were performed using **real world** data from the ad hoc track from TREC, and **simulated** data generated using adapted code from Corsi and Urbano [1].
- TREC data resulted in 177650 pairs of rankings:
  - 110300 pairs were graded on scale [0,1,2,3], represented with **orange** in the scatter plots.
  - 67350 pairs were graded on scale [0,1,2,3,4], represented with **blue** in the scatter plots.
- 2 **highly conjoined simulated** datasets of 10000 ranking pairs were created with a relevance domain of [0,1,2,3,4]
  1. Rankings were generated with a target tau between 0.5 and 1.0 and then **truncated** to lengths between 10 and 100. Relevance scores were assigned based on the **distribution** from 2014 TREC data.
  2. Rankings were generated with target tau between 0.5 and 0.9 and **truncated** to length 100. Relevance scores were **assigned twice**. Once 0 and once 1 was given probability .92, the rest 0.02. Starts with $A_1 = \frac{1}{4}$.
- Test were performed to:
  - Compare the different **normalization factors** (Figure 1-4) and identity-based RBO (Figure 1 & 4).
  - Show the effect of different **maximum** relevance values (Figure 1 & 4).
  - Show the effect of different **dominant** relevance values (Figure 3).
  - Show the effect of different **gain functions** (Figure 4).
- Data points presented are the RBO_{EXT} scores after evaluating up to depth $d = 100$. Persistence was set to $p = 0.9$.

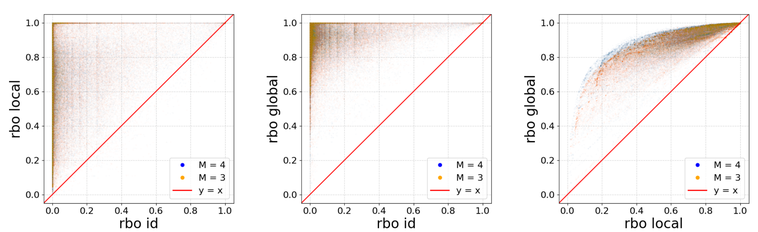

Figure 1: Comparisons between identity-based RBO and the 2 variants for relevance based RBO. TREC data and linear gain used

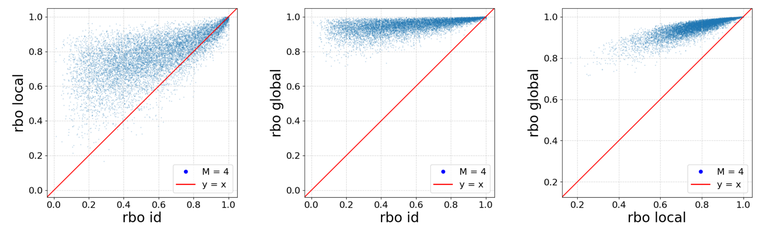

Figure 2: Comparisons between identity-based RBO and the 2 variants for relevance based RBO. Synthetic data and linear gain used

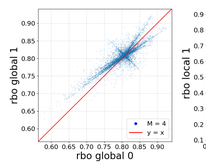

Figure 3: Comparisons between different dominant relevance scores (0,1) using synthetic data and linear gain

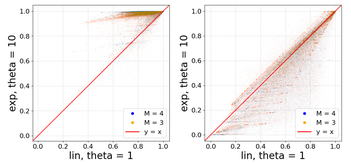

Figure 4: Comparisons between different gain functions for global (left) and local (right) normalization using TREC data

7 - Conclusions

- RBO using **relevance** captures similarity between rankings that **identity-based** RBO does not, as is implied by the scores being **uncorrelated** (Figures 1 & 2).
- The scores from **Global** and **Local** normalization are **correlated** and the score from Global is always **higher** than Local (Figures 1 & 2).
- **Global normalization's** scores tend to **inflate** when the maximum relevance is higher (Figures 1 & 4) and it **captures** similarity between all similar relevance's well (Figure 3).
- **Local normalization** is **less sensitive** to higher maximum relevance (Figures 1 & 4), but **misses** similarity with relevance's of 0 (Figure 3).
- Using **exponential** gain instead of **linear** leads to **higher** scores for Global normalization, while it tends to lead to **lower** scores for Local normalization (Figure 4).

8 - Future work

- Procedure to calculate RBO_{MAX} for **exponential**, or more generally, **arbitrary** gain functions.
- Make the proposed metrics **tie-aware**, like recently done for **identity-based** RBO [1].
- Further analysis of the effect of **changing parameters** such as the gain function, persistence and maximum relevance.
- Explore more ways to define RBO for relevance values, using the **distribution** of relevance scores for instance.

9 - References

[1] Matteo Corsi and Julián Urbano. 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, page 251–260. ACM, July 2024.

[2] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[3] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), November 2010.