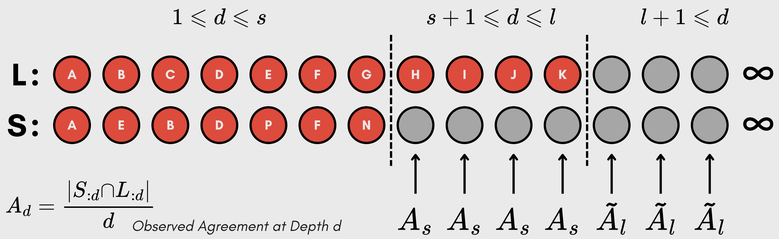
On the Extrapolation of Rank-Biased Overlap and the Assumption of Constant Agreement

Author: Konstantin-Asen Yordanov (k.y.yordanov@student.tudelft.nl)
Supervisor: Matteo Corsi (m.corsi@tudelft.nl)
Responsible Professor: Julián Urbano (j.urbano@tudelft.nl)

PROBLEM DESCRIPTION

- Most Rankings: **top-weighted** (differences in ordering at the top are more significant), **incomplete** (do not cover the entire domain), **indefinite** (evaluation depth is arbitrary)
- Rank-Biased Overlap (RBO): an **overlap-based** and **top-weighted** measure
  - Estimating the **full similarity score** based on an evaluation of the **visible prefixes**
- Persistence Parameter (p): the probability of an arbitrary user continuing to consider the items ranked at the next depth
- Extrapolated RBO: the point estimate that is typically reported and evaluated



First Assumption in Section **[s + 1; l]**

- Assigns the *(l - s)* unseen items in S a **probability of membership** in both rankings by extrapolating the agreement observed at depth *s*

Second Assumption in Section **[l + 1; infinity)**

- Extrapolates the **assumed agreement at depth *l*** across the unseen and potentially infinite tails of the two rankings

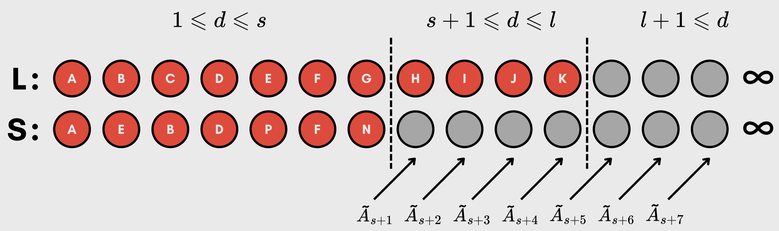
RESEARCH QUESTION

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

- What could serve as a measure of accuracy?
- How could an approach that re-uses the assumed agreement at the previous evaluation depth be implemented?
- How could regression techniques be applied to fit a function on all agreements in the fully-visible section and output a prediction of the agreement at any evaluation depth?

Intuition for Relaxing the Assumptions:

- Interpret agreement as the probability that a randomly-selected element appears in both rankings (i.e. unseen items can be assigned **a degree of fuzzy membership**)
- Estimate that probability to compute assumed agreement **at each depth**, starting at *(s + 1)*



PROPOSED REFORMULATIONS

Techniques for Estimating **Agreement at Rank *k*** as the Probability of an Item Overlapping:

- Previous-Value (**PV**): re-uses the assumed agreement at the previous depth
- Logistic-Regression (**LR**): uses the output of linear-combination-based logistic regression
- Generalized-Additive-Model (**GAM**): uses the output of a non-linear smoothing function

$$\frac{1}{1+e^{-(\beta_0+\beta_1 k)}}$$

Equation for Logistic Regression

$$\frac{1}{1+e^{-[\beta_0+s(k)]}}$$

Equation for GAM

EXPERIMENTAL SETUP AND RESULTS

Data Generation and Testing Configurations:

- 5000 Pairs of Simulated Rankings: number of unique items in the domain chosen as **2000**
- Capturing Incompleteness: pseudo-random generation of *s* and *l*, upper threshold ***l* = 45**
- Varying Persistence: **3 values for *p*** (5, 10, or 20 expected observed items)
- Improving Interpretability: **3 categories of *s*** (small, medium, and large)

Criteria Considered for Performance Evaluation:

- **RBO-Accuracy:** distance between the point estimate and the real RBO score
- **Agreement-Accuracy:** average distance between the assumed and the real agreements from *(s + 1)* to infinity

Table 1: 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 between 0.01 and 0.1, whereas L represents large RBO-distances greater than 0.1.

Type of s	$\lvert RBO - RBO^{OG}_{EXT} \rvert$				$\lvert RBO - RBO^{PV}_{EXT} \rvert$				$\lvert RBO - RBO^{LR}_{EXT} \rvert$				$\lvert RBO - RBO^{GAM}_{EXT} \rvert$			
	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%

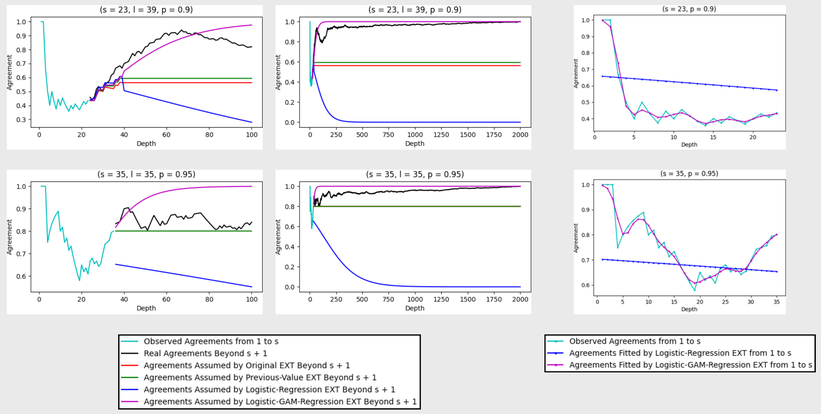


Figure 1: Two instances where GAM outperforms logistic regression in terms of agreement-accuracy. The first two plots show the assumed agreements up to depth 100 (left) and the maximal depth of 2000 (right). The third plot indicates the closeness-of-fit for Logit-Regression and GAM on the observed agreements up to depth *s*. The legends indicate the type and color of each agreement-trace.

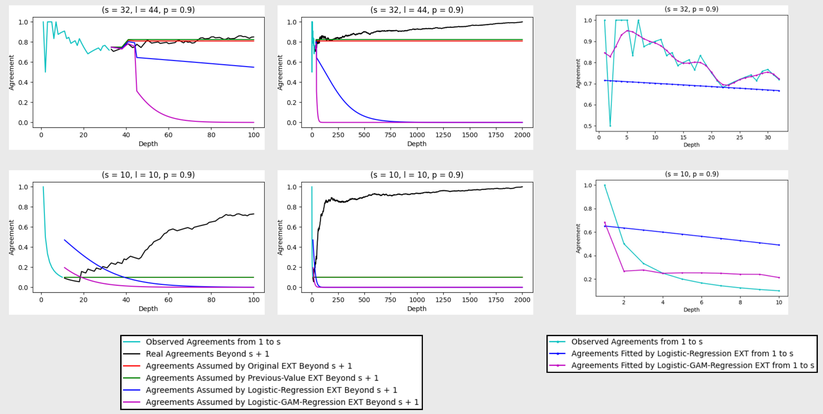


Figure 2: Two instances in which GAM's agreement-accuracy is worse than that of logistic regression. The first two plots show the assumed agreements up to depth 100 (left) and the maximal depth of 2000 (right). The third plot indicates the closeness-of-fit for Logit-Regression and GAM on the observed agreements up to depth *s*. The legends indicate the type and color of each agreement-trace.

CONCLUSIONS AND LIMITATIONS

Findings and Observed Trends:

- Simpler, inflexible approaches (i.e. original and previous-value) retain **constant assumed agreement** beyond depth *l*, without capturing specific patterns in the seen section.
- Logistic regression **fits poorly** during the training phase and performs the worst overall.
- GAM is more flexible, **closely replicating** patterns in the observed agreements.
- The **underfitting-vs.-overfitting** trade-off remains relevant (aiming for a middle-ground).

Important-to-Consider Limitations of the Experiment:

- The simulated rankings are **fully-conjoint**, with an uncharacteristic agreement reaching 1 at the maximal depth of 2000.
- The unseen tails thus **generalize poorly** to realistic scenarios, imposing too strict of a baseline on the measurement of agreement-accuracy.
- A workaround to keep RBO-accuracy unaffected by this limitation is the use of **smaller values for *p*** and the generation of **shorter visible prefixes**.

DIRECTIONS FOR FUTURE WORK

- Use simulation code that is better suited towards RBO and the property of incompleteness
- In the presence of generalizable and informative tails, carry out the experiment using larger values for *p* (in order for top-weightedness to be reduced)
- Redefine extrapolated RBO for other variants of the similarity measure (e.g. tie-handling)

RELATED LITERATURE

- William Webber, Alistair Moffat, and Justin Zobel. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.*, 28(4), 11 2010.
- 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.
- Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297–318, 1986.

EEMCS, Delft University of Technology, CSE3000 Research Project, 2023-2024